hyväksymispäivä arvosana

arvostelija

# Korpusten muodostaminen Web-aineistoista

Juho Kilpikoski

Helsinki 28.10.2010

Seminaariesitys

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

# Sisältö

# 1 Johdanto

Korpukset ovat kielitieteellistä tutkimuskäyttöä varten koostettuja tekstikokoelmia. Korpukseen valitut tekstit pyrkivät muodostamaan edustavan otoksen valitut kriteerit täyttävistä teksteistä. Ne tarjoavat aineistoja, joita vastaan muodostettuja kielitieteellisiä hypoteeseja on helppo testata. Tutkimuskäyttöön tarkoitettuja tekstikokoelmina on ollut olemassa pitkään. Nykyisenkaltainen korpuslingvistinen tutkimus on kuitenkin syntynyt automaattisen tietojenkäsittelyn myötä, sillä vasta sähköisessä muodossa talletetuista korpuksista voidaan tehokkaasti laskea tunnuslukuja. Varhaisimmat yhä käytössä olevat korpukset koostettiinkin 1960-luvulla, kun tietokoneet alkoivat yleistyä yliopistoissa.

World Wide Web on puolestaan 1990-luvun alusta asti kehittynyt kaikille avoin kokoelma Internetissä sijaitsevia hypertekstisivuja. Hyperteksti on seurattavissa olevia, toisiin teksteihin kohdistuvia viittauksia – linkkejä – sisältävää elektronisessa muodossa olevaa tekstiä. Web on kehittyessään luonut kokonaan uudenlaisia tähän hypertekstiluonteeseen perustuvia tekstityyppejä. Esimerkiksi blogikirjoituksissa ja wikisivustoissa toisiin teksteihin linkittäminen on keskeisessä roolissa. Kokonaan uusien tekstityyppien lisäksi monet perinteiset tekstityypit ovat siirtyneet osittain tai kokonaan Webissä esiintyviksi. Webistä onkin kasvanut valtaisa, kaikkien saatavilla oleva kokoelma mitä erilaisempia tekstejä.

Web ei ole jäänyt korpuslingvistikoiltakaan huomaamatta. Heidän mielenkiintonsa tuloksena on kehittynyt kaksi toisistaan selvästi eroavaa tapaa hyödyntää Web korpuksena. Ensimmäinen on kohdella koko olevaa Webiä raakakorpuksena [LEB07, s. 8]. Tällöin esimerkiksi frekvenssejä ei voida laskea suoraan – eihän Webin koko ole tiedossa – vaan niitä pitää arvioida. Nämä arviot perustuvat hakukoneiden antamiin lukuihin löytyneiden tulosten määrästä [MGH03, s. 246].

Toinen tapa on haravoida Webiä ja ladata sopivia Web-sivuja paikallista tarkastelua varten. Tällöin ei tietenkään voida hyödyntää Webin koko laajuutta, mutta toisaalta ladattuja sivuja voidaan tarkastella ja analysoida kuten mitä tahansa raakakorpusta [MGH03, s. 8]. Tämä jälkimmäinen tapa on ainoa järkevä [BUe06, s. 34], ja se toimiikin lähtökohtana tälle esitykselle.

# 2 Korpuksen luominen Web-aineistosta

Korpukset sellaisenaan tarjoavat jonkin verran mielenkiintoista tietoa. Korpuksista voidaan esimerkiksi laskea frekvenssilistoja, joita voidaan suoraan hyödyntää vaikkapa kielen opetuksessa. Tällaiset staattiset listat ja tunnusluvut ovat kuitenkin vain lisäarvo korpusten tärkeimmän tehtävän rinnalla. Kielentutkija voi testata ja kehittää usein subjektiivisten havaintojen pohjalta rakentamaansa hypoteesia korpuksia vastaan, sillä tarjoavathan korpukset hänelle valmiita, kattavia aineistoja todellisesta, käytettävästä kielestä.

Työstettävän hypoteesin tulisi ohjata myös käytettävien aineistojen valintaa: koostetaanhan eri korpukset erilaisten kriteerien perusteella. Tietenkään tarkoituksena ei ole valita vain sellaisia aineistoja joiden tiedetään tai oletetaan tukevan hypoteesia – sehän olisi tulosten vääristelyä! Tavoitteena on päinvastoin karsia pois sellaisia aineistoja, jotka eivät anna tilastollisesti tai muun käytettävän tutkimusmetodologian kannalta merkittäviä tuloksia. Koska Web on hallitsemattoman laaja, tulee juuri korpuksen keräämistä Web-aineistosta rajoittaa työstettävänä olevan hypoteesin perusteella.

## 2.1 Lähdeaineiston valinta ja kerääminen

Aineistona käytettävien Web-sivujen valintaa siis ohjaa voimakkaasti tutkimuskohde ja hypoteesi. Mitä tarkemmin tutkimuskohde on rajattu, sitä kattavammin siihen kuuluvat tekstit voidaan kerätä. Voidaan jopa päästä siihen, että kerättävä aineisto koostuu kaikista Webissä olevista ja kriteerit täyttävistä sivuista. Esimerkiksi Hoffmann on muodostanut varsin laajan transkriptoitua puhetta sisältävän tekstikorpuksen keräämällä kaikki CNN-kanavan tietyllä aikavälillä julkaisemat transkriptit [Hof07].

Tällaisen tarkkarajaisen, muutamissa eri paikoissa sijaitsevan aineiston kerääminen onnistuu yksinkertaisilla komentosarjoilla. Hoffmann on kerännyt aineistonsa Perl-komentosarjoilla, jotka latasivat julkaistut transkriptit osoitteen perusteella [Hof07, s. 70].

Hypoteesi voi kuitenkin olla sellainenkin, että sen tutkimiseen tarvitaan edellistä laaja-alaisempaa aineistoa. Tällöin ei kaikkien aineistojen osoitteita (tai edes osoitteiden määrämuotoa) tunneta ennalta, eikä edellisen kaltaisia yksinkertaisia komentosarjoja voida käyttää. Kuitenkin kerättävää aineistoa voidaan rajoittaa. Esi-

merkiksi yksikielistä korpusta kerättäessä luonnollisena rajaavana tekijänä toimii kieli. Lisäksi laaja-alaisenkin tarkastelun ulkopuolelle on hyvä rajata kielitieteellisesti mielenkiinnottomat sivustot, kuten vaikkapa http-palvelimen antamat virheilmoitukset, jotka toistuvat aina samanlaisina.

Hypoteesi voi lisäksi rajoittaa tarkastelua vaikkapa tekstityypin perusteella. Esimerkiksi blogikirjoituksiin on varsin helppo rajautua, sillä blogit sijaitsevat usein erityisillä blogisivustoilla ja muutenkin seuraavat varsin selkeää ja niille erityistä rakennetta. Lisäksi blogikirjoitukset uutena tekstityyppinä tarjoavat kokonaan uudenlaista aineistoa kielentutkijoille.

Toisin kuin hyvin tarkasti rajatussa tutkimuskohteessa, ei nyt voida tarkastella koko perusjoukkoa, vaan ainoastaan siitä tehtyä otosta. Otoksen tulee olla riittävän laaja jotta sitä voidaan yleistää. Esimerkiksi harvinaista kielen piirrettä tutkittaessa tarvitaan enemmän materiaalia kuin yleistä. Yleisestikin laajemmasta aineistosta saadaan merkitsevämpiä tuloksia kuin pienemmästä. Vaikka suurten aineistojen kerääminen ja käsittely vievätkin enemmän resursseja kuin pienten aineistojen, on yksittäisenkin tutkijan mahdollista muodostaa miljoonien, jopa kymmenien miljoonien sanojen korpuksia. Nämä koot olivat joitakin vuosia sitten vielä hyväksyttäviä suurten perinteisten korpusten kokoja.

Sen miten perusjoukosta kerätään otos, ei pitäisi vaikuttaa tutkimustuloksiin. Siis mikäli samanlaisten valintakriteerien perusteella tehdään kaksi erilaista edustavaa otosta, pitäisi tulostenkin olla samanlaiset.

Laaja-alaisia aineistoja voidaan kerätä Webistä automaattisesti kahdella tavalla: joko hyödyntämällä olemassa olevia hakukoneita sopivien sivujen löytämiseen tai käyttämällä itse hakurobottia (*web crawler*), joka aloittaa haravoinnin sopivilta lähtösivuilta [LEB07, BUe06]. Baroni ja Ueyama pitävät tätä jälkimmäistä lähestymistapaa ainoana järkevänä [BUe06, s. 34].

Ensimmäisessä tavassa hakukoneille syötetään joitakin yleisiä sanoja ja aineistoksi valitaan osa hakukoneen löytämistä sivuista. Nämä sivut sitten ladataan jatkokäsittelyä varten. Tässä tavassa on Baronin ja Ueyaman mukaan samoja ongelmia kuin Webin suorakäytössäkin [BUe06, s. 33], jota käsitellään tarkemmin myöhemmin. Erityisesti hakukoneet eivät näytä kaikkia löytämiään tuloksia. Vaikka hakukone ilmoittaisi tulosten kokonaismääräksi useita miljoonia, niin hakijalle esitetään tyypillisesti alle tuhat tulosta. Tutkijan tarkoituksena on toki tarkastella vain pientä otosta valtavasta perusjoukosta, mutta valinnan pitäisi perustua satunnaisuuteen tai sopivaan, tunnettuun, perusteltuun heuristiikkaan. Hakuko-

neiden heuristiikkoja tutkija ei kuitenkaan voi tietää.

Vaikka Baroni ja Ueyama pitävät oman hakurobotin käyttämistä ainoana järkevänä lähestymistapana, he tunnustavat sen olevan haastavampaa kuin hakukoneisiin turvautuminen. Lähestymistavalla on kuitenkin erittäin merkittäviä etuja: sitä käyttämällä saavutetaan keräysmenetelmän täydellinen hallinta ja mahdollisuus mukauttaa keräysmenetelmää vapaasti [BUe06, s. 34].

Menetelmässä käytetään siis erityistä hakurobottia, jolle annetaan lähtökohdaksi joukko Web-sivuja. Robotti lataa nämä sivut ja etsii sivuilta linkit. Seuraavaksi robotti lataa linkitetyt sivut ja etsii niillä esiintyvät linkit [LEB07, s. 18–19]. Robotti jatkaa tätä rekursiivista keräystä kunnes aineistoja on kerätty riittävästi.

Keräysmenetelmä on hyvin tehokas: selvästikin kohdattavien sivujen määrä kasvaa eksponentiaalisesti rekursioaskeleiden määrään nähden. Edelleen, mikäli menetelmän annettaisiin jatkaa vapaasti, saataisiin kerättyä *kaikki* ne Web-sivut joihin lähtösivulta pääsee mitä tahansa reittiä pitkin. Selvästikin aineistoksi tallennettavia sivuja tulee jotenkin rajoittaa.

Kuten mainittua, hypoteesista nousee luontevasti erilaisia rajoitteita. Joitakin rajoitteista voidaan hyödyntää jo tässä keräysvaiheessa. Esimerkiksi tutkittaessa blogikirjoituksia voidaan rajoittua tarkastelemaan vain linkkejä jotka vievät tunnetuilla blogisivustoilla sijaitseville sivuille.

Hypoteesista nousevien rajoitteiden jälkeenkin voi aineiston määrä jäädä liian suureksi. Tällöin tulee tehdä laadullisten rajoitteiden vielä määrällisiä rajoitteita. Yksinkertaisin määrällinen rajoite on keskeyttää kerääminen manuaalisesti, kun riittäväksi katsottava määrä aineistoja on kerätty. Tällöin rajoitus ei kuitenkaan ole satunnaista vaan on kerätyksi valikoituivat riittävän »lähellä» lähtösivuja olleet sivut.

Toinen tapa on seurata kaikkia linkkejä, mutta tallentaa vain satunnaisesti valittu osa linkitetyistä sivuista – esimerkiksi yksi linkki sadasta kohdatusta. Tällöin todennäköisimmin ladatuksi tulee sellaisia sivuja, joille linkitetään paljon. Tätä voitaneen pitää eräänlaisena käyttöä painottavana heuristiikkana.

Keräämisen lähtökohtana toimivat Web-sivut vaikuttavat suuresti siihen millaisia aineistoja kerätään [LEB07, s. 18–19]. Sopivia lähtösivuja voidaan hakea hypoteesin pohjalta. Esimerkiksi blogisivujen keräämisen sopivia lähdesivuja ovat erilaiset blogilistat ja joukko aktiiviseksi tunnettuja blogeja: blogit tunnetusti linkittävät runsaasti toisiinsa.

## 2.2 Ladatun aineiston käsittely

Viimeistään kun kaikki aineistoksi aiotut sivut on ladattu, tulee varmistaa, että kaikki aineistolle asetetut rajoitteet ovat voimassa. Miten tämä tehdään riippuu tietenkin asetetuista rajoitteista. Tyypillinen rajoite on esimerkiksi vain yhden kielen tutkiminen. Vaikka aineistoa koottaessa olisi pyritty rajoittumaan vain tutkittavalla kielellä oleviin sivuihin esimerkiksi lähtösivujen tai verkko-osoitteiden perusteella, niin väistämättä mukaan eksyy muunkielisiä sivuja. Automaattisen kielentunnistuksen keinoin voidaan lataamisen jälkeen varmistaa, että kaikki korpukseen otettavat tekstit on varmasti kirjoitettu tutkittavalla kielellä.

Webille on tyypillistä, että sama sivu löytyy useammasta kuin yhdestä osoitteesta. Koska emme halua, että tällaiset kaksoiskappaleet vääristävät tuloksia, tulee ne poistaa [BUe06, s. 32]. Tämä onnistuu esimerkiksi laskemalla kaikille ladatuille sivuille kryptografiset tiivistearvot ja poistamalla duplikaatteja siten että kustakin tiivistearvosta huomioidaan vain yksi sivu [BUe06, s. 37].

Kun on varmistettu, että ladatut tiedostot vastaavat rajoitteita eivätkä sisällä kaksoiskappaleita, koossa on suuri joukko tyypillisesti HTML-muodossa olevia tiedostoja. Web-sivuilla on varsinaisten, aineistona kiinnostavien, tekstien lisäksi tyypillisesti myös muuta tekstuaalista sisältöä. Varsinkin kaupallisilla sivustoilla jopa suurin osa sivusta voi olla täytetty toissijaisella materiaalilla. Tämä toissijainen materiaali toistuu usein lähes saman sisältöisenä sivuston eri sivuilla ja voi olla myös kielitieteellisesti toisarvoista.

Korpuksen muodostamista varten onkin ladatut Web-sivut siivottava siivottava niin tästä toissijaisesta sisällöstä kuin teknisistä HTML-merkkauksista. Näistä ensimmäinen on selvästi haastavampaa. Jos aineisto on koottu vain muutamista sivustoista voi olla mahdollista hyödyntää sivuston rakenteeseen perustuvaa pääsisällön tunnistamista. Kaupalliset Web-sivustot tyypillisesti rakennetaan dynaamisesti tietyn mallineen perusteelle ja tällöin sisältö on rakenteellisesti samassa paikassa [Hof07, s. 71]. Kun mallineen rakenne tunnistetaan, voidaan sisältö erottaa tunnistetun rakenteen perusteella vaikkapa jäsennetystä sivusta XPath-lausekkeella.

Yleisessä tapauksessa ei sivujen rakennetta tunneta ennalta, vaan keskeinen sisältö tulee tunnistaa heuristisesti. Tällainen heuristiikka voi perustua esimerkiksi siihen kuinka paljon tekstipätkä sisältää HTML-merkkausta suhteessa koko pituuteensa [BUe06, s. 35].

Niin rakenteeseen perustuvaan kuin herustiseenkin keskeisen sisällön tunnistamiseen voidaan helposti yhdistää HTML-merkkauksen pudottaminen: joko hakemalla jäsennetystä HTML-dokumenttipuusta kaikki tekstisolmut tai primitiivisemmin säännöllisillä lausekkeilla.

Nyt on luotu Web-aineistosta raakakorpus: kokoelma pelkästään luonnollista tekstiä sisältäviä tiedostoja. Tämä voi olla hypoteesin testaamiseksi riittävä aineisto: voidaanhan siitä laskea esimerkiksi frekvenssitietoja. Aineistolle voi olla tarpeen suorittaa tarkempaa analyysiä, jota varten aineisto voidaan esimerkiksi jäsentää morfologisesti. Tällaiset automaattiset jäsennykset ovat kuitenkin samanlaisia kuin perinteisille raakakorpuksille suoritettavat ja siten tämän esityksen ulkopuolella.

## 2.3 Epävarmuustekijöiden huomioiminen

Huomattava on ettei mikään havaintoihin perustuva tieteellinen tutkimus ole täydellistä. Tällaiseen tutkimukseen liittyy aina epävarmuustekijöitä. Tämä on täysin hyväksyttävää. Siispä tarkoituksena ei ole vähätellä epävarmuustekijöitä tai ohittaa niitä ikään kuin niitä ei olisi olemassa. Päin vastoin epävarmuustekijöiden selostaminen ja analysointi lisää tutkimuksen luotettavuutta osoittamalla että tutkimuksen tekijä on pohtinut näitä kysymyksiä.

Tämä pätee myös Web-korpuksia koottaessa. Monet yllä kuvatuista vaiheista ovat heuristisia ja sisältävät useita epävarmuustekijöitä. Tutkijan pitääkin kuvatessaan miten Web-korpuksensa on kerännyt perustella miten nämä epävarmuustekijät on huomioitu. Edellä käytettiin rajausesimerkkinä rajoittautumista vain blogisivustoilla sijaitseviin sivuihin. Kaikki blogit eivät kuitenkaan sijaitse blogisivustoilla. Tällaista rajausta voitaneen kuitenkin pitää riittävänä, varsinkin kun pelkästään blogisivustoiltakin kerättäessä saadaan laaja aineisto.

Tärkeimpien epävarmuustekijöiden selvittämiseksi on hyvä ennen varsinaista Web-korpuksen keräämistä tehdä koekeräys. Tällöin ei tarkoituksena ole saada lopullisia tuloksia, vaan kartoittaa mitä erityisiä ongelmakohtia ja epävarmuustekijöitä kyseisen korpuksen keräämiseen liittyy. Näin havaittuihin kohtiin voidaankin sitten kiinnittää erityistä huomiota varsinaista tutkimusta tehtäessä.

# 3 Web-korpus vs. perinteinen korpus

Web-korpus ei perimmäiseltä olemukseltaan eroa perinteisestä korpuksesta. Molemmat ovat kielitieteellisten hypoteesien testaamista varten koostettuja tekstikokoelmia. Kuitenkin näillä korpustyypeillä on omat tyypilliset piirteensä ja ominaisuutensa.

## 3.1 Mitä lisäarvoa Web tuo perinteiseen korpukseen nähden?

Pohdittaessa Webin tuomaa lisäarvoa kielentutkimukselle mainitaan usein ensimmäisenä sen valtava koko [BUe06, Fle07, Fle10]. Webin kokoa on mahdoton laskea tarkkaan jo senkin takia, että Web kasvaa edelleen jatkuvasti. Hakukoneiden indeksoimien sivujen määrä liikkuu kuitenkin useissa kymmenissä, jopa sadoissa miljardeissa *sivuissa* [Fle10]. Webin todellinen koko voi olla tätä huomattavastikin suurempi. Suurtenkin perinteisten korpusten koot liikkuvat sadoissa miljoonissa *sanoissa*. Selvästikin Web tarjoaa selvästi enemmän raaka-ainetta kielentutkimukseen kuin perinteiset korpukset.

Pelkkä sivujen määrä ei kuitenkaan vielä riittävästi kuvaa Webin laajuutta. Web on sisällöltään uskomattoman rikas. Sen ansiosta, että kuka tahansa voi lukea ja kirjoitta Web-sivuja, tekstilajitkin kattanevat koko kirjoitetun tekstin kirjon. Webistä löytyy niin henkilökohtaisia päiväkirjamaisia tekstejä, kuin kaupallisia tuoteylistyksiä. Tällaisten uusien itsetuotettujen aineistojen lisäksi Webistä löytyy runsaasti myös vanhempia tekstejä, jotka nekin vaihtelevat laidasta laitaan.

Jotkin Webissä ovat kokonaan uusia, eikä niitä ole olemassa Webin ulkopuolella [Fle07, s. 27]. Esimerkiksi wikit, blogit ja keskustelufoorumit ovat Webin omia tekstityyppejä. Näitä tekstityyppejä ei esiinny muissa medioissa, ja mikäli niitä haluaa tutkia ovat Web-korpukset ainoa mahdollisuus.

Webin rikkaudesta ja laajuudesta kertoo hyvin sekin, että jotkut sen teksteistä on kirjoitettu pienillä tai harvinaisilla kielillä. Tällaisille kielille ei ole koskaan koottukaan perinteistä korpusta [Fle07, s. 27]. Toisaalta kaikille suuremmillekaan kielille, kuten esimerkiksi italialle tai japanille ei ole olemassa vakiintunutta referenssikorpusta [BUe06, s. 32] Web on siis tuonut kokonaan uusia kieliä korpuspohjaisen kielentutkimuksen piiriin.

Webiä on pidettyä äärimmäisimpänä monitorikorpuksena [MGH03, s. 242], sillä muuttuuhan se koko ajan: uusia sivuja kirjoitetaan ja vanhoja sivuja katoaa. Tä-

mä tuoreus onkin yksi Webin vahvuuksista [Fle07, s. 27]. Koska Webissä julkaiseminen tapahtuu heti, uudet Web-sivut kuvaavatkin juuri ilmestymishetkellään vallinnutta kielen käyttöä. Kielessä tapahtuvien muutosten pitäisikin siis olla nopeasti havaittavissa Webissä.

## 3.2 Perinteisten korpusten vahvuudet

Yksi perinteisten korpusten suurimmista vahvuuksista on käytön helppous: jos tutkijalla on käyttöoikeus perinteiseen korpukseen, voi hän suoraan hakea vahvistusta hypoteesilleen. Tämä mahdollistaa nopeamman, suorastaan kokeilevan tutkimuksen. Ideoita voi kokeilla heti, ilman että aikaa kuluisi korpusaineiston muodostamiseen.

Käytettäessä tunnettua, vakiintunutta perinteistä korpusta, myös muut tutkijat tuntevat käytetyn korpuksen. Lukiessaan tieteellisestä artikkelista kuinka hypoteesin tarkastelussa on käytetty aineistona *Brittish National Corpus*ta, voi lukija helposti arvioida onko kyseinen korpus sovelias hypoteesin tutkimiseen. Mikäli lukija epäilee korpuksen antamia tuloksia, hän voi jopa testata esitettyä hypoteesia samaista korpusta vastaan. Web-korpusta käytettäessä pitää selostaa koko korpuksen luontiprosessi mahdollisimman tarkasti. Muutoin lukija ei välttämättä vakuutu käytetyn aineiston laadusta, eikä siihen perustuvista tuloksista. Tällaiset selostukset vievät tilaa usein pituudeltaan rajatuissa artikkeleissa varsinaisilta tuloksilta.

Perinteisten korpusten tekijän- ja käyttöoikeudet on myös selvitetty. Tutkija tietään perinteistä korpusta hyödyntäessään täsmälleen mihin hän saa aineistoa käyttää. Webistä kerätyn aineiston oikeudellinen asema on ongelmallinen [Fle10]. Perustuuhan koko Web-korpuksen teko siihen, että Webistä ladataan tekijäoikeuksien suojaamia teoksia, joita lataaja sitten hyödyntää omia tarkoitusperiä varten.

Periteisissä korpuksissa metatiedot voivat olla kattavammat ja erityisesti luotettavammat kuin Web-aineistossa. Perinteisiin korpuksiin valikoituvista teksteistä on tyypillisesti tiedossa vähintään tekijä ja julkaisuaika. Jos nämä tiedot ovat käytettävissä varsinaisessa korpuksessa, voi niihin perustuen tehdä demografista erottelua. Webissä puolestaan kuka tahansa voi itsenään, nimettömästi tai jonain toisena ja julkaisuajankohta voi puuttua tai olla virheellinen. Vähintäänkin Web-aineiston metatietoihin tulee suhtautua suurella varauksella.

## 3.3 Entäpä Webin suorakäyttö?

Sen lisäksi, että Webiä käytetään aineistona korpusten muodostamista varten, voidaan sitä käyttää myös suoraan. Tällöin käytettävän »korpuksen» muodostaa koko Web ja tästä aineistosta käsitellään hakukoneiden avulla [LEB07, s. 8]. Saatavat tulokset ovat siis hakukoneiden antamia osumien määriä sopiville sanoille ja rakenteille [MGH03, s. 246].

Vaikka ensinäkemältä tällainen lähestymistapa vaikuttaa helpolta, niin pian törmätään lukuisiin käytännön ongelmiin [BUe06, s. 33]. Hakukoneiden kyselyissä käytettävä syntaksi on hyvin rajoittunut. Hakukoneet lisäksi normalisoivat kyselyjä: arvaat virheellisiä kirjoituksia, taivuttavat sanoja. Tällainen tiedonhakijalle hyödyllinen toiminta ei suinkaan ole kielitieteilijän mieleen.

Nämä ongelmat ovat kuitenkin pieniä verrattuna siihen, että käyttämällä hakukoneita vaarannetaan koko tutkimuksen tieteellisyys. Tieteellisen tutkimuksen ehkäpä tärkein vaihe – tulosten kerääminen – ulkoistetaan voittoa tavoitteleville yrityksille. Hakukoneiden käyttämät indeksointi- ja järjestysalgoritmit ovat salaisia. Tällaisten salaisten menetelmien käyttäminen on tietenkin avoimen tieteellisen tutkimuksen periaatteita vastaan.

Hakukoneet voivat lisäksi milloin tahansa muuttaa algoritmejaan tai vaihtaa niitä kokonaan toisiin. Kukaan ei takaa, että kysely jolla sai mielenkiintoisa tuloksia tänään, toimisi vielä huomenna. Tästä on ollut jo käytännön tapauskin: Vuonna 2004 AltaVista poisti hakusyntaksistaan `NEAR`-operaattorin. Eräs julkaistu, käytössä ollut tutkimusalgoritmi, joka käytti tätä operaattoria, muuttui näin hyödyttömäksi [BUe06, s. 33].

Hakukoneiden antamat tulokset vaihtelevat huomattavasti. Sama haku voi peräkkäisinä päivinä tuottaa yli kymmenen prosenttia eroavia tuloksia [Kil07, s. 148]. Näin nopeat muutokset vaikuttavat epätodennäköisiltä vaikka kuinka huomioitaisiin Webin nopea päivittymistahti. Edelleen muuttamalla asetuksia, joilla ei pitäisi olla vaikutusta suhteellisiin frekvensseihin, saadaan erilaisia tuloksia [Kil07, s. 150]. Kun alkuperäisen tutkijankin on arvottava mitkä tuloksista olisivat ne luotettavimman, on vaikeaa puhua minkäänlaisesta toistettavuudesta.

Koska ongelmat Webin suorakäytössä hakukoneiden avulla ovat näin perustavaa laatua olevia, ei sitä voida pitää soveliaana välineenä tieteellistä tutkimusta tehtäessä.

# 4 Johtopäätökset

Web-korpusten käyttämisestä voidaankin vetää kaksijakoinen johtopäätös. Ensiksikin on olemassa tutkimuskysymyksiä, joiden tarkasteluun Web-korpukset ovat ainoa mahdollisuus. Toisaalta laajoja aineistoja vaativissa kysymyksissä Web-korpukset täydentävät perinteisiä korpuksia.

Web-keskeisiin tutkimuskysymyksiin Web-korpusten kerääminen on tietenkin enemmänkin kuin perusteltua: Se mahdollistaa kokonaan uudenlaisten tekstien tutkimisen ja avaa näin kokonaan uudenlaisia kielenpiirteitä korpuspohjaisen kielentutkimuksen piiriin.

Web-aineistoista voidaan lisäksi luoda eräänlaisia »täsmäkorpuksia». Kun tutkimuskysymys on kovin tarkkaan rajattu, voivat perinteiset laaja-alaiset korpukset olla liian yleisiä. Niistä ei tällöin voida laskea tilastollisesti merkittäviä tuloksia. Webistä haettavaa aineistoa voidaan kohdentaa tutkimuskysymystä varten, ja saada tarkkaan rajattuihinkin kysymyksiin merkittäviä tuloksia.

Edelleen Webin tuoreus mahdollistaa kokonaan uusien kielen piirteiden seuraamisen lähes reaaliajassa. Toistamalla täsmälleen saman Web-korpuksen muodostusprosessin, on tutkijalla käytettävissään korpuksia, jotka eroavat toisistaan vain keräysajan perusteella. Kun tutkija on saanut keräysprosessinsa automatisoitua, voidaan keräys lisäksi toistaa miten usein tahansa.

Vastattaessa yleisiä aineistoja vaativiin tutkimuskysymyksiin, täydentävät Web-korpukset perinteisiä korpuksia. Vertailtuaan erilaisia aineistoja ja lähestymistapoja oman hypoteesinsa testaamiseen, toteavat Lüdeling *et al.*

> For many linguistic research questions [..] there is no perfect corpus at the moment. [LEB07, s. 15]

Hyödyntämällä sekä perinteistä korpusta, että Web-korpusta voidaan samalla hyödyntää kummankin parhaita puolia: perinteisen korpuksen luotettavuutta sekä Web-korpusten laaja-alaisuutta ja tuoreutta. Parhaimmillaan näin voidaan tällaisiin kysymyksiin saada tuloksia, jotka ovat parempia kuin mitkään yksittäisiin korpuksiin perustuvat tulokset. Lisäksi kun tuloksia on useammista erilaisista aineistoista, voidaan niihin perustuvia johtopäätöksiä pitää yleistettävämpinä.

# 5 Yhteenveto

Web-korpuksilla on puolia, joita perinteiset korpukset eivät voi korvata. Webille omien tekstityyppien tarkasteluun ne tietenkin ovat ainoa mahdollisuus. Lisäksi Web-korpukset ovat tuoreita. Niillä voidaan tarkastella sellaisia kielen piirteitä, jotka ovat kehittyneet tuoreimpienkin korpusten kokoamisen jälkeen.

Luonnollisesti Web-korpuksiin liittyy ongelmia. Webille tyypillinen epävarmuus luotettavuudesta kiusaa heijastuu myös Web-korpuksiin. Erityisesti Web-korpuksiin liittyvää metadataa ei voida pitää läheskään yhtä luotettavana, kuin perinteisten korpusten metadataa. Epäluottamus ulottuu myös itse Web-korpuksiin. Käytettäessä aineisona vakiintunutta perinteistä korpukseen riittää viittaus kyseiseen korpukseen sekä selostus menetelmistä ja tuloksista. Web-korpuksissa puolestaan pitää koko korpuksen muodostamisprosessi esitellä ja perustella, jotta aineistoa voidaan pitää luotettavana.

Web-korpuksia kannattaa käyttää silloin kun se tutkimuskysymysten perusteella Web-aineisto vaikuttaisi tuovan lisäarvoa. Erityisesti tutkittaessa uusia, vain Webissä esiintyviä tekstityyppejä, ovat Web-korpukset ainoa mahdollisuus. Toisaalta Web mahdollistaa aineisojen kohdentamisen juuri tutkimushypoteesin kannalta oleellisiin aineistoihin, mikä voi oleellisesti helpottaa tarkkarajaisten kysymysten tutkintaa. Mikäli aineistoksi tarvitaan laaja-alaisia korpuksia, on Web-korpusten ohella hyvä käyttää perinteisiä korpuksia mahdollisuuksien mukaan.

# Lähteet


BUe06 Baroni, M. ja Ueyama, M., Building general- and special-purpose corpora by web crawling. *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, 2006, sivut 31–40.

Fle07 Fletcher, W. H., Concordancing the web: Promise and problems, tools and techniques. Teoksessa *Corpus Linguistics and the Web*, Hundt, M., Nesselhauf, N. ja Biewer, C., toimittajat, numero 59 sarjassa Languages and Computers, Rodopi, Amsterdam, 2007, sivut 25–45.

Fle10 Fletcher, W. H., Corpus analysis of the world wide web. Julkaisematon käsikirjoitus, 2010.

Hof07 Hoffmann, S., From web page to mega-corpus: the CNN transcripts. Teoksessa *Corpus Linguistics and the Web*, Hundt, M., Nesselhauf, N. ja Biewer, C., toimittajat, numero 59 sarjassa Languages and Computers, Rodopi, Amsterdam, 2007, sivut 69–85.

Kil07 Kilgarriff, A., Googleology is bad science. *Computational Linguistics*, 33,1(2007), sivut 147–151.

LEB07 Lüdeling, A., Evert, S. ja Baroni, M., Using web data for linguistic purposes. Teoksessa *Corpus Linguistics and the Web*, Hundt, M., Nesselhauf, N. ja Biewer, C., toimittajat, numero 59 sarjassa Languages and Computers, Rodopi, Amsterdam, 2007, sivut 7–24.

MGH03 Meyer, C. F., Grabowski, R., Han, H.-Y., Mantzouranis, K. ja Moses, S., The world wide web as linguistic corpus. Teoksessa *Corpus Analysis – Language Structure and Language Use*, Leistyna, P. ja Meyer, C. F., toimittajat, numero 46 sarjassa Languages and Computers, Rodopi, Amsterdam, 2003, sivut 241–254.