*Kimmo Kettunen**

Sanoja analysoivat ja tuottavat ohjelmat hakutermien vaihtelun hallinnassa tekstitiedonhaussa

Tietokoneen avulla tehtävällä tekstitiedonhaulla jossain muodossa on melko tarkalleen 50 vuoden historia. Ensimmäiset tekstitiedonhakujärjestelmät olivat erilaisia viitehakujärjestelmiä, joista haettiin kirjallisuusviitteitä ja tieteellisiä artikkeleita. Hakuperusteina toimivat yksittäiset sanat: tekijöiden nimet, artikkeleita kuvaavat hakusanat jne. Kielenä haussa oli englanti, koska tieteen kieli oli ja on englanti. Hakuohjelmia käyttivät joko tiedonhaun ammattilaiset tai asiaan perehtyneet tutkijat itse. Tiedonhaun yllä leijui silloiselle tietojenkäsittelylle ominainen juhlava tunnelma: se oli toimintaa, jota kuka tahansa ei kyennyt tekemään.

Näistä lähtökohdista tekstitiedonhakuun muodostui pitkä perinne, jossa englanninkielisten dokumenttien tiedonhaku on ollut paljolti tutkimuskohteena aina 1990-luvulle. Kun tekstitiedonhaun kielellistä kattavuutta haluttiin parantaa, englannin sanojen melko yksinkertainen muoto-oppi eli morfologia sai kehittäjät tekemään tietokoneohjelmia, joilla hakutermien muodon vaihtelun vaikutusta dokumenttien löydettävyyteen pyrittiin vähentämään. Ensimmäinen käytetty keino oli hakutermin katkaiseminen sellaisesta kohtaa, että samalla katkaistulla merkkijonolla kyettiin hakemaan useita sanoja. Koska menetelmä kuitenkin on epävarma jättäessään hakutermien katkaisemisen ongelman käyttäjälle, päädyttiin 1960-luvun lopulla siihen, että hakutermien ja teksteissä esiintyvien sanojen yhdenmukaistamiseen on tehtävä oma ohjelma. Tietotekniikka otettiin tässä siis käyttöön myös kielen sananmuotojen käsittelyyn, se ei ollut enää pelkkä dokumenttien varastoinnin, etsimisen ja tulostamisen väline.

Ensimmäisen julkaistun tiedonhaun käyttöön tarkoitetun sananmuotoja yhtenäistävän ohjelman teki Julie Beth Lovins vuonna 1968 Massachusetts Institute of Technologyssa Yhdysvalloissa. Ohjelma tunnetaan alan kirjallisuudessa edelleen Lovinsin karsintaohjelmana eli stemmerinä. Se käsitteli englannin kielen sanoja lopusta päin ja pyrki poistamaan niistä taivutus- ja johdospäätteitä korkeintaan yhden sanaa kohti. Tuloksena oli muodoltaan yhdenmukaistettuja sanoja, joiden avulla dokumenttien indeksin ja käyttäjän antamien hakutermien samankaltaisuutta lisättiin. Ohjelma avasi uuden suunnan automaattiselle sananmuotojen käsittelylle tiedonhaun apuna, ja sittemmin Lovinsin karsintaohjelman tyyppisiä ohjelmia kehitettiin runsaasti 1970- ja 80-luvuilla englantia varten. Tunnetuimmaksi näistä ohjelmista on tullut Martin Porterin vuonna 1980 julkaisema karsintaohjelma, joka tunnetaan myös tekijänsä nimellä Porterin karsintaohjelmana.

Vaikka englannin kielen sanojen muoto-opilliset ongelmat ovat varsin vähäisiä, tiedonhaussa esiteltiin erityisesti 1970- ja 1980-luvuilla monia uusia englannin kielen sananmuotoja käsitteleviä karsintaohjelmia. Niitä tehtiin hiukan eri päämääriin ja eri tavoin, mutta suuria eroja niiden toiminnassa ei ollut. 1980-luvulla voimallisesti kehittynyt kieliteknologia tuotti kuitenkin myös edistyneempiä sana- ja lausetason ohjelmia, joilla voitiin käsitellä kieltä automaattisesti yhä paremmin. Esimerkiksi sanamuotojen täysi automaattinen tunnistus kehittyi 1980-luvulla jokseenkin täydelliseksi kieliteknologiassa. Kesti kuitenkin yllättävän pitkään ennen kuin sanojen perusmuotojen tunnistusta alettiin käyttää yleisemmin tiedonhaun apuna. Suomessa näin tehtiin heti 1980-luvulla, mutta muualla ilmeisesti englannin kielen yleisyys esti näkemästä sen, mitä hyötyä karsintaa kehittyneemmästä tekniikasta olisi tiedonhaussa.

*Kimmo Kettusen väitöskirja *Reductive and generative approaches to morphological variation of keywords in monolingual information retrieval* tarkastettiin Tampereen yliopistossa 27.10.2007.

1990-luvun alussa tiedonhaussa alettiin kiinnostua myös muista kielistä kuin englanti. Yksi ensimmäisistä tutkituista kielistä, joissa oli sanojen muodollista vaihtelua paljon englantia enemmän, oli slovenia, jonka tiedonhausta julkaistiin tutkimustuloksia vuonna 1992. Sittemmin tiedonhakutuloksia on julkaistu kymmenistä kielistä, joista monet ovat muoto-opiltaan vähintään kohtalaisen mutkikkaita. Tällaisia kieliä ovat muun muassa amhara, arabia, heprea, ruotsi, saksa, suomi, turkki, venäjä jne. Kaikkia näitä kieliä varten on tehty erilaisia karsintaohjelmia tiedonhaun käyttöön. Karsinnasta onkin muodostunut viimeistään 1990-luvulla tiedonhaussa hakutermien ja dokumenttiindeksien kielellisen yhtenäistämisen käytännön standardi.

1990-luvulle asti tiedonhaku oli ollut paljolti asiantuntijoiden työtä. Hakukieliin, termistöihin ja aihealueisiin perehtyneet informaatikot tekivät tiedon tarvitsijoille hakuja erilaisista viite- ja kokotekstitietokannoista. Ennen 1990-luvun puoliväliä nopeasti kehittynyt maailmanlaajuinen tietoverkko, World Wide Web, muutti kuitenkin melko nopeasti tilanteen täysin toiseksi: tietoverkkoon ilmestyi ensin aihehakemistoja, sittemmin mahdollistui koko verkon laajuinen yleinen haku hakukoneilla. Viimeisten kymmenen vuoden aikana tekstitiedonhausta on viimeistään tullut jokamiehen laji. Ihmiset tekevät yleisillä hakukoneilla, kuten Google tai Yahoo, jopa päivittäin hakuja joko työhönsä tai harrastuksiinsa liittyen, eikä tätä pidetä sen kummempana taitona. Myös yritysten omista sisäisistä tietoverkoista haetaan tekstitietoa yhä enemmän. Hitaasti mutta selvästi on tapahtunut myös yleisen tietoverkon tietosisältöjen kielellinen muutos: 1990-luvun lopulla arvioitiin seitsemästäkymmenestä kahdeksaankymmentä prosenttia julkisen tietoverkon sisällöstä olevan englanninkielistä. Vuonna 2000 englanninkielisen sisällön määrä oli pudonnut noin 2/3:een, ja vuonna 2002 määrä oli vain hieman yli puolet. Samanlainen muutos on tapahtunut myös tietoverkon käyttäjien kielellisissä taustoissa. Vuonna 2000 hieman yli puolet internetin käyttäjistä oli englannin kielen puhujia, mutta jo vuonna 2004 reilu 2/3 verkon käyttäjistä oli muita kuin englannin puhujia. Selvää onkin, että muun kuin englanninkielisen sisällön määrä verkossa tulee lisääntymään ja internetin käyttäjien kielellinen tausta monipuolistumaan edelleen. Myös niin sanottujen pienten kielten osuus tietoverkossa tulee kasvamaan. Tämä merkitsee sitä, että tietoverkossa on ja tulee olemaan eri kielillä yhä enemmän tekstidokumentteja, joita etsitään hakukoneilla. Tästä selkeästä suuntauksesta huolimatta suurten verkkohakukoneiden kielelliset kyvyt ovat edelleen varsin puutteellisia muiden kuin englannin kielen suhteen. Siksi erityisesti pienten kielten edustajien on syytä tutkia kieltensä tiedonhakua itse, jotta sovellettavaa tietoa olisi tarjolla, jos soveltajia löytyy. Vanhaa fraasia mukaillen voisikin sanoa, että hakukoneiden kielellisten kykyjen kehittäminen on liian tärkeä asia jätettäväksi pelkästään hakukoneyhtiöiden varaan.

Vaikka tekstitiedonhaun reunaehdot ovat muuttuneet alan historian aikana, perusasiat ovat edelleen lähes samat: tekstidokumentteja etsitään antamalla hakutermejä hakuohjelmalle. Hakuohjelman kyky löytää dokumentteja riippuu osin siitä, minkä kielisiä dokumentteja etsitään ja miten hakutermien ja dokumenteissa esiintyvien sanojen muodon vaihtelua käsitellään haussa. Tutkin itse väitöskirjassani sananmuotoja tuottavien ja analysoivien ohjelmien käyttöä hakutermien muodon vaihtelun hallinnassa tekstitiedonhaussa. Tutkimukseni pääkohde on ollut suomen kielen tekstitiedonhaku, mutta esitän tutkimuksessa myös ruotsin, saksan ja venäjän kielen tiedonhakutuloksia. Suomea on yleisesti pidetty vaikeana kielenä tekstitiedonhaussa kielen morfologisen rikkauden vuoksi. Suomen erilaisten sananmuotojen määrän runsaus tekee kyselyissä ja dokumenteissa esiintyvien sananmuotojen täsmäyttämisen ja siten dokumenttien löytymisen normaalia hankalammaksi. Vertailin tutkimuksessani eri menetelmiä tämän ongelman ratkaisemiseksi ja kehitin myös uusia tapoja lähestyä aihetta.

Jaan väitöskirjassani hakutermejä käsittelevät ohjelmat kahteen ryhmään: sananmuotoja analysoiviin eli reduktiivisiin ja sananmuotoja tuottaviin eli generatiivisiin menetelmiin. Sananmuotoja analysoivia menetelmiä ovat karsinta ja perusmuotoistaminen eli lemmaus. Sananmuotoja tuottavia menetelmiä ovat hakuvartaloiden tuottaminen ja niiden kehitelmät sekä täysien hakusanan taipuneiden muotojen tuottaminen.

Osoitan väitöskirjassani, että perusmuotoistamisen lisäksi myös karsinta, taivutusvartaloiden tuottaminen ja siihen perustuvat kehitelmät sekä FCG-menetelmä antavat vähintään kohtalaisia tai hyviä tuloksia suomenkielisessä tekstitiedonhaussa, kun verrokkina käytetään

lemmauksella saavutettavia toistaiseksi parhaita mahdollisia tuloksia. Väitöskirjani kokeelliset tulokset tuovatkin lisää vaihtoehtoja tekstitiedonhaussa käytettäville hakutermin vaihtelun käsittelyn menetelmille suomen kielessä ja muissa morfologisesti mutkikkaammissa kielissä.

Karsinta ja lemmaus ovat kieliteknologisesti sananmuotoja analysoivia ja muotoja johonkin yhteen muotoon palauttavia tekniikoita. Hakutermin erilaisten vaihtelevien muotojen tuottaminen on toinen mahdollinen tapa käsitellä hakutermien muodon vaihtelua. Yksi tutkimukseni keskeisistä tuloksista on, että myös sananmuotoja tuottavat ohjelmat soveltuvat morfologisesti mutkikkaiden kielten hakutermien vaihtelun käsittelyyn nykyisissä enimmäkseen niin sanottuun osittaistäsmäytykseen perustuvissa tekstitiedonhakuympäristöissä. Suomenkielisessä tekstihaussa evaluoin eli arvioin systemaattisesti ensin hakuvartaloiden ja niiden kehitelmien käyttöä. Tämän jälkeen kehitin työssäni uuden menetelmän, FCG:n, morfologisesti vähintään jonkin verran mutkikkaille kielille. Menetelmän keskeinen ajatus on käyttää hakutermeinä annetuista substantiiveista ja adjektiiveista vain niiden tilastollisesti keskeisiä taipuneita muotoja haussa. Menetelmä evaluoitiin neljän kielen tekstitiedonhaussa. Kolmella kielistä, suomella, ruotsilla ja saksalla, menetelmä tuotti hyviä hakutuloksia, kun lemmausta käytettiin verrokkimenetelmänä. FCG-menetelmä perustuu taipuneiden sananmuotojen tilastollisesti vinoihin jakaumiin kielissä, ja siten menetelmä soveltunee myös muihin morfologisesti mutkikkaisiin kieliin.

Olen tehnyt tutkimukseni niin sanotun tiedonhaun laboratoriomallin mukaisessa testiympäristössä suomen ja muiden kielten tunnetuilla testikokoelmilla. Tällaisessa tutkimuksessa ei ole mukana varsinaista oikeaa tiedontarvitsijaa eli hakujärjestelmän käyttäjää. Nykyisten verkkotiedonhakujärjestelmien kattamiin tekstimassoihin verrattuna myös käytettävät testikokoelmat ovat pienehköjä, joskin tilastolliseen tarkastelutapaan riittäviä. Tällaisen tutkimusotteen etuna on kuitenkin se, että tutkittavasta ilmiöstä, tässä tapauksessa hakutermien muodollisen vaihtelun käsittelyn vaikutuksesta tekstitiedonhaussa, saadaan systemaattista tietoa, jota voidaan mahdollisesti soveltaa. Oman tutkimukseni tulosten soveltaminen esimerkiksi verkkotiedonhaun kehittämisessä morfologisesti mutkikkaille kielille on mahdollista, mutta vaatisi myös lisätutkimusta verkkotiedonhakuympäristöissä.

Tieteessä sanotaan kehityksen usein perustuvan siihen, että tutkijat sekä jatkavat aiempaa tutkimusta että kyseenalaistavat entisiä tutkimustuloksia tai vanhoja uskomuksia. Suomessa on oltu tiedonhaun kielellisten kykyjen kehittämisessä valppaina jo yli 20 vuotta. 1980-luvulla kehitettyjen kieliteknologisten ohjelmien käyttöä tiedonhaun apuna alettiin tutkia heti vuosikymmenen puolessa välissä ja tutkimus jatkui aktiivisena 1990-luvun ajan. Tähän historiaan on ollut sekä hyvä asettua että hyödyntää sen tuloksia, mutta toisaalta myös yrittää nähdä ja tehdä asioita hieman toisin.