*Eija Airio**

Tiedonhaun ja kieltenvälisen tiedonhaun morfologiset ongelmat.

Lingvististen metodien ja sumeiden merkkijonotäsmäytysvälineiden soveltaminen

Tiedonhaku perustuu eksakteihin matemaattisiin algoritmeihin. Tietoa haetaan useimmiten luonnollisella kielellä ja haun kohteena ovat luonnollisella kielellä kirjoitetut tekstit. Luonnollinen kieli on kehittynyt ihmisten välisessä vuorovaikutuksessa. Sille on ominaista joustavuus ja tietynlainen epämääräisyys. Kieleen voidaan luoda uusia sanoja ilmaisemaan uusia käsitteitä tarpeen tullen. Esimerkiksi tekniikan kehittymisen myötä on ollut tarpeen keksiä nimitykset uusille asioille, kuten *tietokone, tiedosto* tai *kännykkä*. Joskus uudelle ilmiölle löytyy nimitys laajentamalla jo olemassa olevan sanan merkitystä: *verkko* ei enää tarkoita ainoastaan kalaverkkoa. Yhdellä sanalla voi siis olla monia merkityksiä. Toisaalta joskus yhteen ja samaan asiaan voidaan viitata monin eri sanoin: voidaan puhua koirasta, piskistä, hauvasta tai rakista. Kukin näistä sanoista viittaa samaan eläimeen, mutta eri sivumerkityksin. Luonnollinen kieli toimii erinomaisesti ihmisten välisessä kommunikaatiossa. On kuitenkin selvää, että kielen joustavuus on ristiriidassa tiedonhaun eksaktien algoritmien kanssa.

Luonnollinen kieli voidaan jakaa viiteen alijärjestelmään: semantiikkaan, fonologiaan, leksikologiaan, syntaksiin ja morfologiaan. Semantiikka käsittelee sanojen merkitystä, fonologia kielen äännerakennetta ja leksikologia kielen sanoja. Syntaksi puolestaan on lauseiden rakenteen ja lauseenmuodostussääntöjen tutkimusta. Morfologia keskittyy sanatasolle: tarkastellaan esimerkiksi sitä, miten sanat taipuvat ja miten sanoja johdetaan.

Tutkimukseni käsittelee morfologisten ilmiöiden vaikutusta tiedonhakuun. Tutkimukseni kannalta oleellisia käsitteitä ovat yhdyssana, fraasi, sanan perusmuoto ja sanan vartalo. **Yhdyssana** on sana, joka koostuu kahdesta tai useammasta autonomisesta sanasta, jotka on joko kirjoitettu yhteen (esimerkiksi *luentosali*) tai liitetty toisiinsa yhdysviivalla (esimerkiksi *kuorma-auto*). **Fraasi** puolestaan tarkoittaa yhdyssanan vastinetta kielessä, jossa ei yleensä käytetä yhdyssanoja. Esimerkiksi *luentosali* on englanniksi *lecture hall*, erikseen kirjoitettuna. Sanan **perusmuoto** on selvä käsite: se tarkoittaa esimerkiksi substantiivin nominatiivimuotoa, kuten *sali*. **Vartalo eli stemmi** on se sanan osa, joka jää jäljelle, kun taivutuspäätteet ja tunnukset on erotettu. Suomenkielessä sanalla voi olla useita erilaisia vartaloita, kuten vokaalivartalo ja konsonanttivartalo, kun sen sijaan englannissa niitä on vain yksi.

Tiedonhaku perustuu dokumenttien indeksointiin: kukin dokumentin sana tallennetaan indeksiin. Erillisiksi sanoiksi tulkitaan useimmiten välimerkin tai välilyönnin toisistaan erottamat merkkijonot. Sanat voidaan tallentaa siinä muodossaan kuin ne dokumentissa esiintyvät. Jos dokumentissa siis esiintyy sana *luentosalissa*, niin tallennetaan indeksiin *luentosalissa*. Kun tiedonhakija tällöin kirjoittaa hakukenttään sanan *luentosali*, se ei täsmää tuohon indeksin sanaan *luentosalissa*. Tällaista indeksiä, johon sanat tallennetaan taipuneessa muodossa, kutsutaan **taivutusmuotoindeksiksi**. Kun tehdään hakuja systeemissä, joka perustuu taivutusmuotoindeksiin, olisi siis syytä laittaa kyselyyn sanojen kaikki tai ainakin tärkeimmät taivutusmuodot, siis esimerkiksi *luentosali, luentosalit, luentosalissa* jne. Tärkeimmät taivutusmuodot voidaan generoida automaattisesti. Tällaista menetelmää kutsutaan **generatiiviseksi menetelmäksi**.

Reduktiivinen menetelmä toimii päinvastoin kuin generatiivinen menetelmä.

---

*YTM Eija Airion väitöskirja *Morphological Problems in IR and CLIR. Applying linguistic methods and approximate string matching tools* tarkastettiin Tampereen yliopistossa 13.6.2009

Indeksointivaiheessa pyritään löytämään kunkin sanan kaikille taipuneille muodoille yksi yhteinen muoto eli normalisoidaan sanat. Perusmuotoistaminen on eräs normalisointapa: tallennetaan indeksiin sanan perusmuoto. Tällöin indeksointiin pitää liittää kyseisen kielen perusmuotoistaja. Useat perusmuotoistajat pystyvät pilkkomaan yhdyssanan osiinsa. On siis mahdollista, että yhdyssanan lisäksi indeksiin tallennetaan sen osat, siis esimerkiksi kun dokumentissa esiintyy sana *luentosalissa,* tallennetaan indeksiin *luentosali, luento* ja *sali*. Toinen yleinen normalisointimetelmä on stemmaus: se tarkoittaa sitä, että tallennetaan indeksiin kunkin sanan vartalo. Tällöin indeksointiin on liitettävä kyseessä olevan kielen stemmeri.

Kun normalisoidusta indeksistä haetaan tietoa, pitää tietysti käyttää normalisoituja sanamuotoja: perusmuotoja tai vartaloita. Tiedonhakijoita voidaan esimerkiksi ohjeistaa käyttämään sanojen perusmuotoja tai sitten hakusysteemiin voidaan liittää sama stemmeri tai perusmuotoistaja kuin mitä on käytetty indeksointivaiheessa.

Kieltenvälinen tiedonhaku tarkoittaa sitä, että kyselyn kieli ja haettavien dokumenttien tai www-sivujen kieli eroavat toisistaan. Kyselykieltä kutsutaan lähtökieleksi ja haettavien dokumenttien kieltä tai kieliä kohdekieleksi. Kieltenvälinen tiedonhaku on joko kaksikielistä tai monikielistä. Kun kohdekieliä on yksi, on kyseessä kaksikielinen tiedonhaku. Esimerkiksi kirjoitetaan kysely suomeksi, ja haetaan ranskankielisiä dokumentteja tai sivuja. Kaksikielisestä tiedonhausta on hyötyä henkilölle, joka ei osaa muotoilla kohdekielistä, esimerkiksi ranskankielistä, kyselyä, mutta kykenee lukemaan kohdekielellä kirjoittua tekstiä (tai vaihtoehtoisesti saa käännösapua).

Monikielisessä tiedonhaussa kohdekieliä on useita. Kirjoitetaan siis kysely vaikkapa suomeksi, ja haetaan esimerkiksi ranskan-, espanjan-, italian- ja saksankielisiä dokumentteja. Monikielisestä tiedonhausta on hyötyä henkilölle, joka osaa useita kieliä: hänen ei tarvitse tehdä erillistä hakua kullakin kielellä, vaan yksi haku riittää. Monikielistä tiedonhakua voidaan käyttää hyväksi myös esimerkiksi silloin, kun tehdään hakuja kuvatietokannasta, jossa kuvailut on tehty useammalla eri kielellä. Tällöin hakijan ei edes tarvitse osata kohdekieliä, koska haun kohteena ovat kuvat.

Kieltenvälinen tiedonhaku perustuu kääntämiseen: joko kysely käännetään kohdekielelle (tai kohdekielille) tai dokumentit käännetään lähtökielelle. Jälkimmäinen vaihtoehto olisi varmaankin tiedonhakijalle mieluista: hän saisi dokumentit luettavakseen omalla kielellään. Tämä vaihtoehto on kuitenkin kallis ja hankala toteuttaa. Yleisimmin käännetäänkin kysely kohdekielelle tai kohdekielille. Tavallisimmat käännösmenetelmät ovat korpuksiin perustuva kääntäminen, konekääntäminen ja sanakirjaan perustuva kääntäminen. Käytän tutkimuksessani pääosin sanakirjaan perustuvaa kääntämistä, yhdessä testissä myös konekääntämistä. Mikään näistä käännösmenetelmistä ei ole ongelmaton. Sanakirjaan pohjautuvalle kääntämiselle aiheuttavat vaikeuksia kääntymättömät sanat, yhdyssanat, fraasit, sanojen taipuminen ja sanojen monimerkityksisyys.

Monikielinen tiedonhaku on mutkikkaampaa kuin kaksikielinen, koska kohdekieliä on useita. Jos kutakin kohdekieltä vastaa erillinen indeksi, pitää suorittaa haku kustakin indeksistä erikseen ja sitten yhdistää tuloslistat. Jos taas kohdeindeksi on Web-tyyppinen monikielinen indeksi, voidaan vaihtoehtoisesti suorittaa yksi monikielinen haku: siis yhdistetään kaikki käännetyt kyselyt yhdeksi kyselyksi. Tuloslistojen yhdistäminen on näistä yleisempi menettelytapa monikielisen tiedonhaun tutkimuksessa. Osittain tämä johtuu siitä, että monikielisen kyselyn muotoileminen on vaikeaa.

Tiedonhaussa voidaan käyttää avuksi sumeita merkkijonotäsmäytysmenetelmiä. Menetelmistä suosituin on n-grammaus. N-grammi tarkoittaa n:n pituista alkuperäisen merkkijonon alimerkkijonoa. N-grammausta voidaan käyttää esimerkiksi silloin, kun halutaan löytää parhaimmin täsmäävät sanat indeksin sanojen tai sanakirjan sanojen joukosta. Kieltenvälisessä tiedonhaussa n-grammausta voidaan käyttää kääntymättömiin sanoihin: tällainen sana voi olla erisnimi tai erikoistermi, joka ei sisälly sanakirjaan, mutta on kirjoitusasultaan samankaltainen lähtö- ja kohdekielessä.

Tiedonhaun tutkimuksessa on perinteisesti nojattu niinsanottuun laboratoriomalliin. Se tarkoittaa sitä, että oikeat tiedonhakijat on korvattu testikokoelmilla. Testikokoelma koostuu staattisesta dokumenttijoukosta, staattisista hakuaiheista sekä relevanssikorpuksesta. Laboratoriomallin etu on se, että eri tutkijoiden eri aikoina suorittamien testien tuloksia voidaan helposti verrata keskenään. Laboratoriotestit ovat luotettavia myös siksi, ettei tarvitse ottaa huomioon testihenkilöiden ominaisuuksien vaikutusta hakutuloksiin. Laboratoriotestien

avulla voidaan verrata esimerkiksi erilaisten haku- tai indeksointitapojen vaikutusta tulokseen. Niiden perusteella ei kuitenkaan voida päätellä, miten hyödylliseksi tiedonhakijat kokevat jonkin systeemin tai miten hyviä tuloksia jokin menetelmä antaa kun kyselyitä muotoileekin oikea tiedonhakija.

Kieltenvälisessä tiedonhaussa on perinteisesti verrattu kohdekielisen kyselyn ja käännetyn kyselyn tulosta toisiinsa laboratoriotestein. Sekä lähtökielinen kysely että kohdekielinen kysely on tällöin syntyperäisen kielentaitajan muotoilema. Otetaan esimerkiksi englanti-suomi tiedonhaku, ja haetaan tietoa koulujen kesälomista Suomessa. Lähtökielinen kysely voisi olla ”summer holiday school Finland”. Sanalle *summer* käyttämämme sanakirja antaa käännöksen *kesä*. *Holiday* kääntyy sanoiksi *vapaapäivä* ja *loma*. Sanalle *school* löytyy käännökset *koulu, osasto, tiedekunta, yliopisto, korkeakoulu, koulukunta, parvi, koulia* ja *opettaa*. Sana *Finland* kääntyy sanaksi *Suomi*. Englanti-suomi -tiedonhaussa yritetään siis muotoilla suomenkielinen kysely näistä käännöksistä. On melkoisen varmaa, että suomea osaavan henkilö muotoilema kysely, esimerkiksi ”kesäloma koulu suomi” antaa tässä tapauksessa paremman tuloksen kuin käännetty kysely. Testien perusteella onkin todettu, että lähtökielestä käännetty kysely antaa huomattavasti huonomman tuloksen kuin kohdekielinen kysely.

Väitöskirjaani sisältyvistä tutkimuksista neljä perustuu laboratoriomalliin. Yhden tutkimuksen tulokset perustuvat käyttäjätesteihin.

Tutkimuskysymykseni käsittelevät mm. monikielistä tiedonhakua, yhdyssanojen pilkkomisen vaikutusta hakutulokseen tiedonhaussa ja kaksikielisessä tiedonhaussa, kaksikielistä tiedonhakua taivutusmuotoindeksissä, sanakirjan laadun vaikutusta kaksikielisen tiedonhaun tulokseen sekä kaksikielisen tiedonhaun etuja käyttäjille. Esittelen seuraavaksi tutkimukseni oleellisimmat tulokset.

Monikielinen tiedonhaku ja tuloslistojen yhdistäminen: Vertailin neljän erilaisen tuloslistojen yhdistämisalgoritmin vaikutusta hakutulokseen. Lisäksi vertailin erilaisten indeksin normalisointitapojen vaikutusta. Vaikka listojen yhdistämisalgoritmit olivat erilaisia, tulokset eivät juuri eronneet toisistaan: kaikki antoivat melko heikon tuloksen. Sen sijaan indeksin normalisointitavalla oli suuri merkitys: perusmuotoistaminen antoi paremman tuloksen kuin stemmaus. Tulokset ovat yhteneväisiä aiempien monikielisen tiedonhaun tulosten kanssa: hyvää tuloslistojen yhdistämisalgoritmia ei ole löydetty – kaikki antavat melko huonon tuloksen verrattuna kaksikieliseen tiedonhakuun

Yhdyssanojen vaikutus tiedonhakuun ja kaksikieliseen tiedonhakuun: Testasin yhdyssanojen pilkkomisen vaikutusta hakutulokseen suomen-, ruotsin- ja saksankielisessä tiedonhaussa sekä englanti-suomi, englanti-ruotsi ja englanti-saksa -haussa. Tutkimukseni perusteella yhdyssanojen pilkkominen osiinsa indeksointivaiheessa parantaa hakutulosta huomattavasti kaksikielisessä tiedonhaussa, kun lähtökieli on fraasiorientoitunut kieli ja kohdekieli yhdyssanakieli. Tämä johtuu siitä, että fraasin osat käännetään erikseen, mutta indeksissä esiintyy vain kokonainen yhdyssana, ellei yhdyssanoja ole pilkottu. Jos englanti-suomi haussa esiintyy fraasi *lecture hall*, ja käännetään sana kerrallaan, saadaan kohdekieliseen kyselyyn *luento sali*, erikseen kirjoitettuna. Indeksissä sen sijaan on vain luentosali, yhteenkirjoittuna, ellei yhdyssanoja ole pilkottu indeksointivaiheessa. Myös tavallisen, siis yksikielisen tiedonhaun tulokseen yhdyssanojen pilkkominen vaikutti positiivisesti, mutta ei niin paljon kuin kaksikielisen tiedonhaun tulokseen.

Kaksikielinen tiedonhaku taivutusmuotoindeksissä: Tutkimukseni mukaan kaksikielisen tiedonhaun tulos taivutusmuotoindeksissä on huono ainakin silloin, kun kohdekieli on voimakkaasti taipuva kieli. Tämä johtuu siitä, että sanakirja antaa vain sanan perusmuodon, kun taas indeksissä esiintyy sanoja taipuneessa muodossa. Testasin kahden menetelmän, sanamuotojen generoinnin ja n-grammauksen, vaikutusta tiedonhaun tulokseen. Kielipareina olivat englanti-ruotsi, englanti-suomi, ruotsi-suomi ja suomi-ruotsi. Sanamuotoja generoitaessa kullekin sanakirjan antamalle käännökselle siis generoidaan oleellisimmat taivutusmuodot ja liitetään ne kohdekieliseen kyselyyn. N-grammausta puolestaan hyödynnettiin siten, että kullekin käännökselle etsittiin indeksistä parhaimmin täsmäävät sanat, jotka usein (mutta ei aina) olivat sanan taivutusmuotoja. Molemmat menetelmät paransivat hakutulosta huomattavasti verrattuna siihen, että kysely oli muodostettu käännössanakirjan antamista sanoista, jotka ovat useimmiten sanojen perusmuotoja.

Kaksikielisen tiedonhaun hyöty käyttäjille: Käyttäjätesteihin perustuvan tutkimukseni ideana oli selvittää, miten kaksikielinen tiedonhaku toimii oikeassa tiedonhakutilanteessa. Perinteisestihän

tähän on käytetty laboratoriotestejä. Rekrytoin Tampereen yliopiston opiskelijoita testihenkilöiksi. Haut tehtiin Googlella Webissä. Tässäkin suhteessa käyttäjätestini siis erosi laboratoriotutkimuksista, jotka tehdään staattisessa kokoelmassa. Kielipareina tässä testissä olivat suomi-ruotsi, suomi-ranska ja englanti-saksa. Esimerkiksi suomi-ruotsi testissä koehenkilöitä pyydettiin muotoilemaan suomenkielinen kysely, joka sitten käännettiin ruotsiksi, ja lisäksi ruotsinkielinen kysely. Ne koehenkilöt, joiden ruotsinkielen taito oli hyvä, pystyivät muotoilemaan onnistuneita kyselyitä ruotsiksi. Tällöin saatiin samantapaisia tuloksia kuin laboratoriotesteissä: käännetyt kyselyt antavat paljon huonomman tuloksen kuin kohdekieliset kyselyt. Sen sijaan koehenkilöt, joiden ruotsinkielen taito oli kohtalainen tai huono, muotoilivat epäonnistuneita ruotsinkielisiä kyselyitä: he unohtelivat sanoja ja tekivät kirjoitusvirheitä. Näin ollen he saivatkin parempia tuloksia suomesta ruotsiin käännetyillä kyselyillä kuin muotoilemillaan ruotsinkielisillä kyselyillä. Käyttäjätestien tuloksena siis oli, että kaksikielinen tiedonhaku on hyödyllisempää kuin laboratoriotestien perusteella on päätelty, olettaen että sanakirja on laadukas.

Sanakirjan laadun vaikutus kaksikielisen tiedonhaun tulokseen: On selvää, että käännössanakirjan laatu vaikuttaa kaksikielisen tiedonhaun tulokseen. Tutkimukseni perusteella vaikutus on kuitenkin niin suuri, että huonon sanakirjakäännöksen antama hyöty voidaan asettaa kyseenalaiseksi. Käyttäjätestini perusteella voidaan todeta, että huonoon sanakirjaan perustuva systeemi ei auta kieltenvälisessä tiedonhaussa edes henkilöitä, joiden kohdekielen taito on heikko.

Kielellä on suuri vaikutus tiedonhaun ongelmien laatuun ja vaikeuteen. Englanti on perinteisesti ollut yleisin testikieli tiedonhaun tutkimuksissa. Informaatiotutkimuksen ja interaktiivisen median laitoksella on luontevasti englannin lisäksi keskitytty suomeen. Suomi on voimakkaasti taipuva yhdyssanakieli, englanti taas heikosti taipuva fraasikieli. Voimakkaasti taipuva kieli on tiedonhaun kannalta vaikeampi kuin heikosti taipuva kieli ja yhdyssanat aiheuttavat tiedonhaulle enemmän ongelmia kuin fraasit. Suomi ja englanti ovat siis erinomaiset valinnat, kun tutkitaan kielen vaikutusta tiedonhaun tuloksiin.

Lopuksi voisi todeta, että vaikka morfologian vaikutusta tiedonhakuun ja kieltenväliseen tiedonhakuun on tutkittu paljon, paljon on vielä tutkimatta. Voisi olla hyödyllistä suunnata tutkimusta nykyistä enemmän käyttäjätesteihin päin: niiden avulla saataisiin uutta arvokasta tietoa, mitä ei ole mahdollista saavuttaa laboratoriotesteillä.