Anni Järvelin, Sanna Kumpulainen,
Ari Pirkola & Eero Sormunen

Sumeat käännösmenetelmät läheisten sukulaiskielten välisessä tiedonhaussa

Anni Järvelin, Sanna Kumpulainen, Ari Pirkola & Eero Sormunen, Sumeat käännösmenetelmät läheisten sukulaiskielten välisessä tiedonhaussa [Fuzzy translation techniques in cross-language information retrieval between closely related languages] Informaatiotutkimus 25 (4), p. 86-96.

This article presents results from a study, where fuzzy string matching techniques were used as the sole query translation technique in Cross Language Information Retrieval (CLIR) between the closely related Scandinavian languages Swedish and Norwegian. It is a novel research idea to apply only fuzzy string matching techniques in query translation. Closely related languages share a number of words that are cross-lingual spelling variants of each other. These spelling variants can be translated by means of fuzzy matching. When cross-lingual spelling variants form a high enough share of the vocabulary of related languages, the fuzzy matching techniques can perform well enough to replace the conventional dictionary-based query translation. Different fuzzy matching techniques were tested in CLIR between Norwegian and Swedish and it was found that queries translated using s-gram matching and a combined technique of transformation rule based translation (TRT) and n-grams performed well. For the best fuzzy matching query types performance difference with respect to dictionary translation queries was not statistically significant.

Address: Anni Järvelin, Department of Information Studies, FIN-33014 University of Tampere, Finland. E-mail: anni.jarvelin@uta.fi

1. Johdanto

Kieltenvälinen tiedonhaku tarkoittaa tiedonhakua, jossa tiedonhakijan esittämät kyselyt ovat erikielisiä kuin dokumenttikokoelma, josta tietoa haetaan. Kraaij (2004) tarjoaa katsauksen kieltenvälisen tiedonhaun menetelmiin. Kieltenvälinen tiedonhaku perustuu ajatukseen, että monet henkilöt ymmärtävät useita kieliä. Tyypillisesti vieraan kielen tuottaminen on sen ymmärtämistä vaikeampaa ja tämän vuoksi kyselyiden muotoilu vieraalla kielellä saattaa olla työlästä, vaikka tekstin ymmärtäminen olisikin melko ongelmatonta. Haettaessa tietoa laajoista monikielisistä kokoelmista olisikin tiedonhakijan kannalta hyödyllistä voida esittää hakupyyntö järjestelmälle yhdellä kielellä, mutta saada vastaukseksi dokumentteja useammalla kielellä.

Norja ja ruotsi ovat sukulaiskieliä, jotka ovat kehittyneet läheisessä kulttuurisessa ja historiallisessa yhteydessä toisiinsa. Kielten sanastosta yli 90 % on samankaltaista, jos tietyt ortografiset erot ja taivutuspäätteet jätetään huomiotta (Barðdal ym. 1997). Kielten samankaltaisuudesta seuraa, että tekstin ymmärtäminen kielten välillä on melko ongelmatonta: Norjankielinen henkilö pystyy ymmärtämään ruotsinkielistä tekstiä ilman erityistä harjoittelua, ja päinvastoin. Tämän vuoksi ruotsinkieliset aineistot ovat norjankielisille tiedonhakijoille hyödyllisiä, jos ne vain onnistutaan löytämään. Kielten oikeinkirjoituksen eroista johtuen kyselyn muotoileminen voi kuitenkin olla hankala tehtävä tiedonhakijalle, joten sitä ei pitäisi jättää käyttäjän harteille.

Tyypillisesti kieltenvälinen tiedonhaku perustuu joko kyselyiden tai doku-

menttikokoelman kääntämiseen. Käytetyin lähestymistapa on kyselyiden automaattinen kääntäminen dokumenttikokoelman kielelle käännössanakirjojen avulla (Kishida 2005). Sanakirjakäännös on melko tehokas (effective) menetelmä, mutta kärsii sanakirjojen rajallisesta kattavuudesta ja jatkuvasta päivitystarpeesta, mikä nostaa menetelmän kustannuksia. Kyselyiden ja dokumenttikokoelman kielten ollessa läheisiä sukulaiskieliä, jotka jakavat suuren määrän kieltenvälisiä kirjoitusasuvariantteja, eli kirjoitusasultaan ja merkitykseltään samankaltaisia erikielisiä sanoja, tulee yksinkertaisempien sumeiden käännösmenetelmien käyttö mahdolliseksi. Tiedonhakuun läheisten sukulaiskielten välillä ei ole aiemmassa tiedonhauntutkimuksessa kiinnitetty erillistä huomiota. Se on kuitenkin mielenkiintoinen tutkimusalue, sillä kielten läheinen sukulaisuus tekee sanakirjaperustaisen kyselynkäännöksen korvaamisen sitä yksinkertaisemmilla ja halvemmilla sumeilla menetelmillä mahdolliseksi.

Sumeat merkkijonomenetelmät perustuvat oletukseen, jonka mukaan kaksi sanaa joiden merkkijonot ovat samankaltaiset, ovat myös merkitykseltään samankaltaisia (Robertson & Willett 1998). Niitä käytetään merkkijonojen samankaltaisuuden tunnistamiseen ja mittaamiseen, ja ne ovat hyödyllisiä monissa tilanteissa tiedonhaussa. Esimerkiksi sanojen morfologisesta variaatiosta ja kirjoitusvirheistä seuraava sanamuotovariaatio on tietokannoissa yleistä. Saman sanan eri muotojen tunnistaminen on tiedonhaun kannalta tärkeää, sillä ne edustavat samaa käsitettä ja ovat siten samanarvoisia tiedonhakijan esittämien kyselyiden kannalta. Kieltenvälisessä tiedonhaussa sumeita merkkijonomenetelmiä voidaan käyttää esimerkiksi käännössanakirjoista puuttuvien hakuavainten kieltenvälisten kirjoitusasuvarianttien tunnistamiseen dokumenttikokoelman indeksistä (Pirkola ym. 2002; Keskustalo ym. 2003).

N-grammit ovat tiedonhaun sovelluksissa runsaasti käytettyjä sumeita merkkijonomenetelmiä. N-grammi-täsmäytys on todettu tehokkaaksi menetelmäksi muun muassa haettaessa erisnimien väärinkirjoitettuja vastineita tietokannoista (Pfeiffer ym. 1996) ja modernien sanojen historiallisten kirjoitusasuvarianttien tunnistamisessa historiallisia dokumentteja sisältävistä tietokannoista (O'Rourke ym. 1997). McNamee ja Mayfield (2003) ovat myös käyttäneet korpusperustaista n-grammikyselynkäännösmenetelmää kieltenvälisessä tiedonhaussa. Pirkola ym. (2002) ja Keskustalo ym. (2003) tutkivat n-grammi-täsmäytystä ja sen versiota s-grammi-täsmäytystä kieltenvälisessä tiedonhaussa. Menetelmiä käytettiin käännössanakirjoista puuttuvien sanojen kääntämiseen: Sanakirjoista puuttuvat sanat ovat tyypillisesti erisnimiä ja teknisiä erityistermejä, jotka ovat usein kieltenvälisiä kirjoitusasuvariantteja. Tähän perustuen, mahdollisia käännösvastineita pyrittiin tunnistamaan kohdekielisen dokumenttikokoelman indeksistä n- ja s-grammien avulla. Tutkimuksissa todettiin erityisesti s-grammit tehokkaaksi sanakirjakäännöstä täydentäväksi menetelmäksi.

Pirkolan ym. (2003) ja Toivosen ym. (2005) kehittämä tilastollisiin transformaatiosääntöihin perustuva TRT-menetelmä (Transformation Rule-based Translation) on sumea käännösmenetelmä, joka on myös todettu tehokkaaksi käännössanakirjoista puuttuvien kieltenvälisten kirjoitusasuvarianttien käännöksessä. Kahden kielen välillä esiintyvät johdonmukaiset ortografiset erot rekisteröidään automaattisesti perustuen listaan samanmerkityksisiä sanapareja näillä kielillä ja sääntöjä käytetään sitten tiedonhaussa lähdekielisten sanojen muokkaamisessa lähemmäksi kohdekielistä kirjoitusasua. Transformaatiosääntöjä voidaan soveltaa yksinään tai n-grammitäsmäytykseen yhdistettynä. N- ja s-grammit ovat kieliriippumattomia menetelmiä, joita voidaan soveltaa uusien kieliparien käsittelyyn vaivattomasti. Myös tilastolliset uudelleenkirjoitussäännöt ovat automaattisesti luotavissa uusille kielipareille. Menetelmät soveltuvat siten hyvin kieltenväliseen tiedonhakuun: ne ovat halpoja ja yksinkertaisia kyselynkäännösmenetelmiä, jotka eivät vaadi jatkuvaa ylläpitoa. Ne ovat ihanteellisia käännösmenetelmiä läheisten sukulaiskielten välillä, jos kyselynkäännöksen laatu saavuttaa riittävän korkean tason ollakseen kilpailukykyinen sanakirjakäännöksen kanssa.

Tutkimuskysymykset, joihin tässä artikkelissa pyritään vastaamaan, ovat:

1) Ovatko sumeat täsmäytysmenetelmät kilpailukykyisiä kyselynkäännösmenetelmiä verrattuna sanakirjaperustaiseen käännökseen tiedonhaussa lähisukulaiskielten välillä?

2) Mikä tutkituista sumeista käännösmenetelmistä soveltuu parhaiten kyselynkäännökseen norjan ja ruotsin välillä?

Luvussa 2 keskustellaan lyhyesti ruotsin ja norjan kielen ominaisuuksista. Luku 3 esittelee tutkimuksessa käytetyt sumeat käännösmenetelmät ja luku 4 käytetyt tutkimusmenetelmät ja -aineistot. Luvussa 5 esitellään tutkimuksen tulokset ja luvussa 6 yhteenveto ja johtopäätökset.

2. Ruotsin ja norjan kielen ominaisuuksista

Norja on ruotsin kieltä läheisimmin muistuttava kieli ääntämykseltään, taivutusmorfologialtaan, sanavarastoltaan ja syntaksiltaan. Norjassa on käytössä kaksi laillisesti tasa-arvoista kirjakieltä, bokmål ja nynorska. Bokmålin juuret ovat tanskan kielessä, josta se on lukuisten uudistusten kautta kehittynyt ja jota se edelleen läheisesti muistuttaa. Nynorska kehitettiin vastareaktiona bokmålin samankaltaisuudelle tanskan kanssa ja sen suurelle erolle kansan puhekielestä. (Nationalencyklopedi 1990, osa 14, s. 287.) Norjalaisista noin 80 % käyttää ensisijaisesti bokmålia, joka on siis kielistä selvästi yleisempi ja myös tässä tutkimuksessa käytetty norjan kieli. Jatkossa ilmaisulla norjan kieli viitataankin (hieman harhaanjohtavasti) nimenomaan bokmåliin.

Barðdalin ym. (1997, 128-129) tutkimuksen mukaan noin 91 % norjan (bokmål) ja ruotsin kielen sanastosta on keskenään identtistä tai samankaltaista. Norjan ja ruotsi ovat ortografialtaan varsin samankaltaisia kieliä, mutta joitakin johdonmukaisia eroja kielten välillä esiintyy: Kielten aakkostot ovat muuten samat, mutta ä ja ö kirjoitetaan norjan kielessä tanskalaiseen tapaan æ ja ø. C:n, x:n ja z:n käyttöä vältetään norjan kielessä, myös vierasperäisissä sanoissa (senter ”center”, sjakk ”schack”). Ruotsin kielessä käytettävät päätteet tion/-sion/-ssion kirjoitetaan norjan kielessä –sjon (stasjon, misjon, pensjon). Pitkien konsonanttien kaksoiskirjoittamisessa on osittain erilaiset säännöt kuin ruotsin kielessä ja norjassa kirjoitetaan usein h ennen v-kirjainta, kuten sanassa hvit. D-kirjain on karsittu useista sanoista, niin että sanan kirjoitusasu vastaa yleisintä ääntämystä: li ”lid”, skei ”sked”, kunne ”kunde”. Sen sijaan d säilyy norjan kielessä adjektiivien taivutusmuodoissa: god – godt. Yksi tärkeimmistä eroista on vanhojen diftongien säilyminen norjan kielessä ruotsia laajemmassa mittakaavassa (esim. røyk ”rök”). (Nationalencyklopedin 1990, osa 14, 285.)

Sekä ruotsin että norjan kielessä on melko rikas taivutusmorfologia. Molemmissa kielissä substantiivit taipuvat suvun ja sijamuotojen mukaan, sekä määräisessä ja epämääräisessä muodossa. Substantiivien suku vaikuttaa myös niiden yhteydessä esiintyvien adjektiivien muotoon. Taulukossa 1 esitetään esimerkkejä substantiivien ja adjektiivien taipumisesta. Ruotsin kielessä substantiiveilla on kaksi sukua, uter ja neuter. Norjan kielessä on käytössä kolmen suvun järjestelmä: uter-sukuiset substantiivit jaetaan norjan kielessä maskuliineiksi ja feminiineiksi. (Barðdal ym. 1997, 285-289.) Ruotsin kielessä substantiivit jaetaan viiteen deklinaatioon sen mukaan millaisella suffiksilla niiden monikkomuodot muodostetaan (-or, -ar, -er, -n, -). Norjan kielessä substantiivien monikkomuodot ovat yksinkertaisempia: uter-sukuiset substantiivit saavat monikossa aina päätteen –er ja neuter-sukuiset substantiivit joko päätteen –er tai ei päätettä lainkaan. Molemmissa kielissä on käytössä kaksi sijamuotoa, nominatiivi ja genetiivi. Genetiivimuotoa ilmaistaan päätteellä –s. (Barðdal ym. 1997, 293-301.) Määräistä artikkelia tai demonstratiivipronominia seuraavat substantiivit ovat molemmissa kielissä määräisessä muodossa (Nationalencyklopedin 1990, osa 14, s. 285).

Adjektiivit taipuvat pääsanansa suvun ja luvun lisäksi vertailumuodoissa. Ruotsin kielessä komparatiivi ja superlatiivi muodostetaan päätteillä -are ja -ast (vit, vitare, vitast), norjan kielessä päätteillä -ere ja -est (hvit, hvitere, hvitest). Osa adjektiivien komparatiivimuodoista on epäsäännöllisiä molemmissa kielissä. (Barðdal ym. 1997, 307-308.) Verbit taipuvat aikamuodoissa ja kahdessa tapaluokassa, infinitiivissä ja imperatiivissa. Infinitiivin pääte on ruotsissa -a ja norjan kielessä -e (Nationalencyklopedin 1990, osa 14, 285). Verbit jaetaan vahvoihin ja heikkoihin verbeihin molemmissa kielissä. Heikot verbit taipuvat aikamuodoissa säännönmukaisesti, vahvojen verbien taivutuksessa tapahtuu muun muassa vokaalimuunnoksia. Ruotsin kielessä mahdollisia heikkojen verbien imperfektin päätteitä ovat -ade, -dde, -te ja -de (esimerkiksi köpa, köpte), bokmålissa -et/-a, -dde, -te ja -de (esimerkiksi kjøpe, kjøpte). (Barðdal ym. 1997, 318-331.)

Sekä ruotsin että norjan kielessä esiintyy runsaasti yhdyssanoja ja johdoksia. Yhdyssanat kirjoitetaan yhteen ja yhdyssanan pääsana on

Taulukko1. Substantiivien taivutusmuotoja ruotsin ja norjan kielessä, sekä adjektiivien taipuminen pääsanansa mukaan.

substantiivin muoto	**ruotsi**	**norja**
en / ei(n)	vit häst	hvit häst
ett / eit	vitt hus	hvitt hus
tre	vita hästar; vita hus	hvite hester / hvite hus
den; det; de	vita hästen; vita huset; vita husen	hvite hesten; hvite huset; hvite husene

yleensä sen jälkimmäinen osa. Merkitykseltään leksikalisoituneet yhdyssanat ovat yleisiä (jordgubbe). Yhdyssanojen muodostamiseen käytetään usein fogemorfeemeja, joiden käyttö on kuitenkin melko epäsäännöllistä. Esimerkkejä fogemorfeemeista ja niiden käytöstä ovat muun muassa vokaalin poistaminen (skola – skolhus), vokaalin vaihtaminen (flicka – flickebarn) ja -s:n lisääminen (utvikling – utviklingsarbeit). Johdosten muodostamisessa käytetään norjan ja ruotsin kielessä tyypillisesti samankaltaisia suffikseja. Esimerkiksi ruotsin kielessä tyypillistä antonyymien muodostukseen käytettävää prefiksiä -o, vastaa usein norjankielessä -u (olycklig – ulykkelig).

3. Sumea käännös

3.1 N- ja s-grammit

N-grammi-täsmäytys on sumea merkkijonomenetelmä, jonka avulla voidaan tunnistaa ja mitata merkkijonojen samankaltaisuutta. Kieltenvälisen tiedonhaun kannalta menetelmä toimii siten, että käännettävät hakuavaimet ja kohdekielisen dokumenttikokoelman indeksin sanat pilkotaan n-grammeiksi, eli lyhyemmiksi merkkijonoiksi, joiden pituus on n. Hakuavainten ja indeksitermien samankaltaisuus lasketaan vertailemalla niistä muodostettujen n-grammien joukkoja perustuen merkkijonojen yhteisten n-grammien ja toisistaan eroavien n-grammien lukumäärään (ks. esimerkiksi Pirkola ym. 2002 laskukaavaa varten). Tavallisesti käytettyjä n-grammeja ovat muun muassa kahden merkin mittaiset digrammit (n=2) ja kolmen merkin mittaiset trigrammit (n=3). Hakuavaimen käännökseksi valitaan sitä läheisimmin vastaavat kohdekielisen indeksin merkkijonot. Robertson ja Willett (1998) esittävät katsauksen n-grammien sovelluksiin.

N-grammit muodostetaan useimmiten merkkijonojen vierekkäisistä merkeistä. Pirkola ym. 2002) kehittivät uuden n-grammi-täsmäytysmenetelmän, jossa n-grammeja muodostaessa voidaan hypätä joidenkin merkkien yli. Menetelmän nimi, s-grammi-menetelmä, viittaakin sanaan skip, hypätä yli. Perinteiset n-grammit voidaan nähdä tämän menetelmän erikoistapauksena, jossa hypätään nollan merkin yli. S-grammit ryhmitellään luokkiin perustuen ylihypättyjen merkkien määrään ja vain samaan luokkaan kuuluvia s-grammeja vertaillaan keskenään. S-grammiluokka ilmaisee siis kuinka monen merkin yli s-grammeja muodostettaessa hypätään. Merkkiyhdistelmäindeksi (CCI, Character Combination Index) ilmaisee kaikkien merkkijonosta muodostettavien s-grammiluokkien joukkoa (Keskustalo ym. 2003). Esimerkiksi CCI{{0},{1,2}} ilmaisee, että merkkijonosta muodostetaan kaksi s-grammiluokkaa: yksi, jossa muodostetaan perinteisiä n-grammeja vierekkäisistä merkeistä ja toinen, jossa s-grammit muodostetaan sekä hyppäämällä yhden että hyppäämällä kahden merkin yli. Taulukossa 2 esitetään esimerkkejä sekä perinteisistä n-grammeista, että s-grammeista. Erimittaisia s-grammeja voidaan muodostaa samalla tavalla kuin n-grammeja. Nimenomaan s-digrammit ovat kuitenkin soveltuvia kieltenvälisen kirjoitusasuvariaation mallintamiseen.

3.2 TRT

Tilastollisiin transformaatiosääntöihin perustuva käännös (TRT, transformation rule based translation) on Pirkolan ym. (2003) kehittämä sumea käännösmenetelmä, joka perustuu tilastollisesti tuotettujen uudelleenkirjoitussääntö

Taulukko 2. Esimerkkejä s-grammeista sanalle abradacabra. N-grammit ovat s-grammeja, joiden CCI={0}.

Tyyppi	CCI	Muodostettavat s-grammit
digram	{0}	{ab, br, ra, ad, da, ac, ca, ab, br, ra}
trigram	{0}	{abr, bra, rad, ada, dac, aca, cab, abr, bra}
s-digram	{1}	{ar, ba, rd, aa, dc, aa, cb, ar, ba}
s-digram	{2}	{aa, bd, ra, ac, da, ab, cr, aa}
s-digram	{{0},{1}}	{{ab, br, ra, ad, da, ac, ca, ab, br, ra}, {ar, ba, rd, aa, dc, aa, cb, ar, ba}}
s-digram	{{0},{1, 2}}	{{ab, br, ra, ad, da, ac, ca, ab, br, ra}, {ar, ba, rd, aa, dc, cb, ar, bd, ra, ac, da, ab, cr}}

jen käyttöön käännettäessä kirjoitusasuvariantteja kieleltä toiselle. Transformaatiosäännöt kuvaavat samaan kirjoitusjärjestelmään kuuluvien kielten oikeinkirjoituksessa esiintyviä säännönmukaisia eroja. Kieltenväliset kirjoitusasuvariantit eroavat toisistaan usein säännönmukaisin tavoin. Esimerkiksi merkkijono *for* norjankielisen sanan alussa muuttuu usein merkkijonoksi *för* ruotsinkielisen sanan alussa (*forbund – förbund*). Transformaatiosääntöjen avulla lähdekielisiä sanoja voidaan muuttaa kohdekielisiksi vastinsanoikseen tai muokata lähemmäksi niitä.

Transformaatiosääntöihin perustuvaa käännösmenetelmää voidaan soveltaa yksinään tai n-grammeihin yhdistettynä, jolloin lähdekielinen sana muokataan ensin lähemmäksi kohdekielistä kirjoitusasua transformaatiosääntöjen avulla. Tämän jälkeen saatu käännösvastine voidaan täsmäyttää n-grammien avulla kohdekieliseen indeksiin oikean käännösvastineen löytämiseksi. Transformaatiosääntöihin perustuvan käännöksen ideaa ja transformaatiosääntöjen tuottamista käsitellään laajemmin artikkeleissa Pirkola ym. (2003); Toivonen ym. (2005).

Transformaatiosääntö sisältää ne lähde- ja kohdekieliset kirjaimet, joita sääntö koskee, sekä niitä ympäröivät merkit. Lisäksi jokaisen säännön yhteyteen tallennetaan säännön esiintymisfrekvenssi lähdeaineistossa, jota sääntöjen luomiseen käytetään, sekä luotettavuuskerroin. Säännön luotettavuuskerroin lasketaan jakamalla säännön frekvenssi niiden lähdesanojen lukumäärällä, joissa säännön lähdekielen merkkijono esiintyy. Frekvenssiä ja luotettavuuskerrointa voidaan käyttää luotettavimpien sääntöjen valitsemiseksi käännöstä tehtäessä. Esimerkki säännöstä, jolla muokataan norjankielisen sanan kirjoitusasua lähemmäksi ruotsinkieltä, on:

for för beginning 132 147 89.80

Sääntö luetaan: kirjain o ennen kirjainta r ja kirjaimen f jälkeen, muutetaan kirjaimeksi ö kun merkkijono esiintyy sanan alussa, säännön luotettavuuskertoimen ollessa 89.80 %. Kerroin lasketaan säännön frekvenssistä (132) ja niiden sanojen lukumäärästä lähdeaineistossa, joissa merkkijono for esiintyy (147).

4. Tutkimusmenetelmät ja aineistot

4.1 Hakuaiheet ja testikokoelma

Sumeisiin käännösmenetelmiin perustuvien kyselyiden tehokkuuden tutkimiseksi tehtiin laboratoriotutkimus, jossa 60 norjankielistä hakutehtävää käännettiin ruotsiksi useilla eri sumeilla käännösmenetelmillä. Tutkimuksessa käytettiin vuoden 2003 CLEF (Cross-Language Evaluation Forum) tutkimusympäristöä, joka sisältää 60 kappaletta hakutehtäviä sekä dokumenttikokoelman ja relevanssiarviot useille kielille, mukaan lukien ruotsi (Peters 2003). Tutkimusympäristö ei sisällä norjankielisiä hakuaiheita, minkä vuoksi englanninkieliset hakuaiheet käännettiin norjaksi. Käännöstyön teki äidinkielenään norjaa puhuva henkilö. Hakuaiheiden sanat perusmuotoistettiin käyttäen Lingsoft Oy:n morfologiaohjelmia Swetwol ja Nobtwol. Yhdyssanat ositettiin ja niiden osat normalisoitiin. Sulkusanat poistettiin normalisoinnin jälkeen. Kymmeneen hakuaiheeseen liittyvät kyselyt epäonnistuivat alustavissa testeissä johtuen teknisistä ongelmista. Nämä hakuaiheet jätettiin pois lopullisista testeistä ja lopulliset testiajot suoritettiin jäljelle jääneillä 50 hakuaiheella.

Testikokoelmana käytettiin ruotsinkielistä CLEF testikokoelmaa, joka sisältää 142819 dokumenttia ja on kooltaan 352 Mb. Dokumentit ovat ruotsalaisen uutistoimiston Tidningarnas Telegrambyrån (TT) vuosina 1994-1995 julkaisemia uutissähkeitä (Peters 2003). Testikokoelman sanat perusmuotoistettiin käyttäen Lingsoft Oy:n morfologiaohjelmaa Swetwolia. Yhdyssanat ositettiin ja sekä alkuperäinen yhdyssana että sen osat perusmuotoistettiin ja indeksoitiin. Sanat, joita morfologiaohjelma ei tunnistanut, indeksoitiin sellaisinaan erilliseen indeksiin ja indeksi jaettiin siten kahteen osaan: tunnistettujen- ja tunnistamattomien sanojen indeksiin. Tutkimuksessa käytettiin probabilistista InQuery tiedonhakujärjestelmää, joka tarjoaa monipuolisia operaattoreita erilaisten kyselyrakenteiden muodostamiseen (Broglio ym. 1993).

4.2 TRT sääntöjen luominen

TRT-sääntöjen tuottamiseen tarvitaan laajahko vastinsanaparilista lähdekielisille sanoille ja niiden kohdekielisille vastineille. Sanaparilista norjan- ja ruotsinkielisille sanoille muodostettiin kääntämällä osa ruotsinkielisen tutkimuskokoelman indeksistä norjaksi GlobalDix sanakirjan avulla. Tunnistamattomat sanat poistettiin, jolloin jäljelle jäi 6714 norja-ruotsi sanaparia sisältävä vastinsanaparilista. Sanaparit, joiden edit distance arvo oli suurempi kuin puolet sanaparin pidemmän sanan pituudesta, poistettiin. Samoin poistettiin sanaparit joihin sisältyi alle neljän kirjaimen mittainen sana. Lopulliseen sanaparilistaan sisältyi 3058 sanaparia. Tämä osoittautui liian pieneksi lähdeaineistoksi sääntöjen luomiseen: sääntöjen frekvenssi jäi niin alhaiseksi, että hyvien ja huonojen sääntöjen erotteleminen toisistaan automaattisesti oli hankalaa. Käännöksessä käytettiin transformaatiosääntöjä, joiden luotettavuuskerroin oli vähintään 50 % ja frekvenssi vähintään 2. Nämä kynnysarvot rajasivat käytettävissä olevien sääntöjen määrän melko pieneksi.

4.3 Kyselyt

N-grammien, s-grammien ja transformaatiosääntöihin perustuvan käännöksen suorituskykyä kyselynkäännösmenetelminä testattiin viiden testikyselytyypin avulla. N-grammeista testattiin digrammeja (*n-gram*). S-grammeja testattiin kahdella eri merkkiyhdistelmäindeksillä: CCI:n ollessa {{0},{1}} (*Skip1*) ja CCI:n ollessa {{0},{1, 2}} (*Skip2*). Transformaatiosääntöihin perustuvaa käännöstä testattiin sekä yksinään (*TRT*) että yhdistettynä perinteisiin digrammeihin (*TRT-n-gram*). N- ja s-grammi-kyselyihin valittiin jokaiselle lähtökieliselle sanalle neljä käännösvastinetta. Kaksi käännöstä valittiin tunnistettujen sanojen indeksistä n- tai s-grammi-käännöksen tuloslistan kärjestä ja kaksi käännöstä tunnistamattomien sanojen indeksistä. Valinta perustui Hedlundin ym. (2004) tutkimukseen, jossa todettiin n-grammi-hakuavainten lisäämisen kyselyyn heikentävän kyselyn tarkkuutta siten, että paras keskitarkkuus saavutettiin sisällyttämällä kyselyyn vain muutamia n-grammi-hakuavaimia.

TRT-kyselyyn otettiin mukaan kaikki transformaatiosääntöjen tuottamat käännösvastineet. Käännösvastineita saatiin tuotettua matalista kynnysarvoista huolimatta melko vähän; kun n- ja s-grammi kyselyissä oli keskimäärin 36,9 hakuavainta (9,2 avainta * 4 käännösvastinetta), oli TRT-kyselyissä keskimäärin vain 13,9 hakuavainta, siis vain hieman enemmän kuin yksikielisissä norjankielisissä kyselyissä (9,2 avainta). Monille hakuavaimille ei saatu tuotettua lainkaan käännösvastineita TRT-sääntöjen avulla, jolloin TRT-kyselyyn jäi hakuavaimeksi pelkästään alkuperäinen norjankielinen hakuavain. Kun transformaatiosääntöihin perustuva käännös yhdistettiin n-grammien käyttöön, valittiin n-grammi-täsmäytykseen vain paras transformaatiosääntöjen tuottama käännösvastine. Neljä parasta n-grammi käännösvastinetta valittiin kyselyyn, kuten muissakin n- ja s-grammi-kyselyissä. Kaikissa sumein menetelmin käännetyissä kyselyissä käytettiin synonyymirakennetta, jolloin yhden sanan kaikki käännösvastineet yhdistettiin toisiinsa InQueryn synonyymi-operaattorin avulla. Tämä niin sanottu Pirkolan menetelmä (Pirkola method, Pirkola 1998) on todettu aiemmissa tutkimuksissa toimivaksi kyselyrakenteeksi kieltenvälisessä tiedonhaussa.

Näitä kyselyitä verrattiin kolmeen vertailukohtakyselyyn: kahteen kääntämättömään kyselyyn, ruotsinkieliseen (*swebase*) ja norjankieliseen (*nobase*), sekä sanakirjan avulla käännettyyn kyselyyn (*dicbase*). Sanakirjaperustainen kyselynkäännös oli tärkein vertailukohta, sillä se on vakiinnuttanut asemansa

yleisimpänä kieltenvälisessä tiedonhaussa käytettävänä kyselynkäännösmenetelmänä. Norjankieliset kyselyt käännettiin manuaalisesti ruotsiksi GlobalDix-sanakirjan avulla. Yhdyssanat ositettiin ja kaikki sanakirjan lähdesanoille ehdottamat käännökset otettiin mukaan kyselyyn. Sanat, joille ei löytynyt yhtään käännösvastinetta, lisättiin kyselyihin sellaisenaan ja merkittiin tunnistamattomiksi sanoiksi. Yhden sanan kaikki käännösvastineet yhdistettiin toisiinsa InQueryn synonyymi-operaattorilla.Lisäksi käännöksessä saatujen sanaliittojen yhdistämisessä käytettiin InQueryn uwn-läheisyysoperaattoria (unorderd window of length, jossa n:n arvosana 7). Ruotsinkielisistä hakuaiheista muodostettu kysely tarjosi vertailukohdan yksikieliseen tiedonhakuun. Norjankielinen kysely toimi heikkona vertailukohtana, josta nähtiin miten hyvin haku onnistuu ruotsinkielisestä kokoelmasta norjankielisellä kyselyllä. Tästä nähtiin tarvitaanko käännöstä ylipäätään, ja miten paljon sumea käännös parantaa tuloksia. Yksikieliset kyselyt olivat rakenteettomia kyselyitä. Esimerkkejä kyselytyypeistä esitetään liitteessä.

4.4 Suorituskyvyn mittarit

Käännösmenetelmien suorituskykyä vertailtiin perustuen niiden saavuttamaan yhdelletoista saantitasolle laskettuun keskiarvoiseen keskitarkkuuteen (MAP, Mean Average Precision), joka laskettiin yli kaikkien kyselyiden. Lisäksi kyselytyyppien suorituskykyä arvioitiin interpoloituna keskitarkkuutena yhdellätoista vakiolla saantitasoilla, joista piirrettiin saantitarkkuuskäyrä. Tulosten tilastollisen merkitsevyyden laskemiseen käytettiin Friedmanin kaksisuuntaista järjestyslukutestiä. Tilastolliset merkitsevyystasot esitetään tulostaulukoissa.

5. Tulokset

5.1 Sumeat käännösmenetelmät vs. sanakirjakäännös

Tutkimuksen tulokset olivat lupaavia. Sumeat käännösmenetelmät osoittautuivat kilpailukykyisiksi menetelmiksi sanakirjakäännöksen kanssa: paras sumea menetelmä (skip2) saavutti keskimäärin n. 85 % sanakirjakyselyn suorituksesta. Parhaiten menestyivät s-grammeihin perustuvat menetelmät, samalla kun TRT-käännös yksinään ei osoittautunut tässä tutkimuksessa riittäväksi käännösmenetelmäksi. Taulukossa 3 esitetään interpoloimattomat keskiarvotarkkuudet kaikille kyselytyypeille, sekä testikyselyiden ja vertailukohtakyselyiden keskiarvotarkkuuksien prosentuaaliset erot. Koska n-grammien, s-grammien ja yhdistetyn TRT-n-grammimenetelmän tulokset olivat lähellä toisiaan, niihin viitataan tässä luvussa grammikyselyinä. Näiden sumeiden käännösmenetelmien suorituskyvyn eroja tutkitaan luvussa 5.2.

Sanakirjakäännökseen perustuva vertailukohtakysely (dicbase) saavutti menetelmistä korkeimman keskiarvotarkkuuden (34,1 %) ruotsinkielisen kyselyn ollessa toiseksi paras (31,8 %). Grammikyselyiden keskitarkkuudet jäivät noin 16–22 prosenttia sanakirjakyselyä heikommiksi. Erot sanakirjakyselyyn ja ruotsinkieliseen kyselyyn eivät ole tilastollisesti merkitseviä. Sparck Jones (1975) esittää, että kahden menetelmän välinen ero on käytännössä huomattava, jos se on tilastollisesti merkitsevä ja ylittää 5 prosenttiyksikköä ja käytännössä olennainen, jos ero ylittää 10 prosenttiyksikköä. Tämän perusteella grammiperustaisten sumeiden käännösmenetelmien ja sanakirjakäännöksen keskinäinen ero ei käytännössä ole olennainen. Sen sijaan grammikyselyiden suorituskyky on tilastollisesti merkitsevästi norjankielistä vertailukohtakyselyä parempi. Keskitarkkuuksien prosentuaalisen eron ollessa selvästi yli 10 prosenttiyksikköä kaikkien grammikyselyiden ja norjankielisen vertailukohtakyselyn välillä, niiden eron voidaan sanoa olevan myös käytännössä olennainen. Pelkkä TRT-kysely suoriutui norjankielistä kyselyä paremmin saavuttaen 33,5 prosenttia sitä korkeamman keskitarkkuuden. Tämä ero ei ollut tilastollisesti merkitsevä. Tulosta arvioidessa on tärkeää huomioida, että TRT-käännös tuotti vain niukasti käännösvastineita kyselyiden hakuavaimille ja oli siten melko samankaltainen norjankielisen kyselyn kanssa. Sanakirjakysely ja ruotsinkielinen kysely olivat TRT-kyselyä olennaisesti parempia eron ollessa tilastollisesti merkitsevä.

5.2 Sumeiden käännösmenetelmien vertailu

Sumeita menetelmiä vertailtiin myös keskenään kieltenväliseen tiedonhakuun parhaiten soveltuvan

Taulukko 3. Norjasta ruotsiin käännettyjen kyselyiden keskitarkkuusarvot (MAP) sekä sumeiden käännösmenetelmien suorituskyvyn erot verrattuna vertailukohtakyselyihin (%), N=50. (* tilastollisesti merkitsevä ero, ** tilastollisesti erittäin merkitsevä ero)

Kysely	Nobase	Swebase	Dicbase	Skip1	Skip2	N-gram	TRT	TRT-n-gram
Keskitarkkuus	12,6	31,8	34,1	28,3	28,6	26,5	16,9	27,7
Ero Nobaseen	0	+151,3	+170	+124,2*	+126,5*	+109,9**	+33,5	+119,5**
Ero Swebaseen		0	+7,5	-10,8	-9,9	-16,5	-46,9**	-12,7
Ero Dicbaseen			0	-17,0	-16,1	-22,3	-50,5**	-18,7

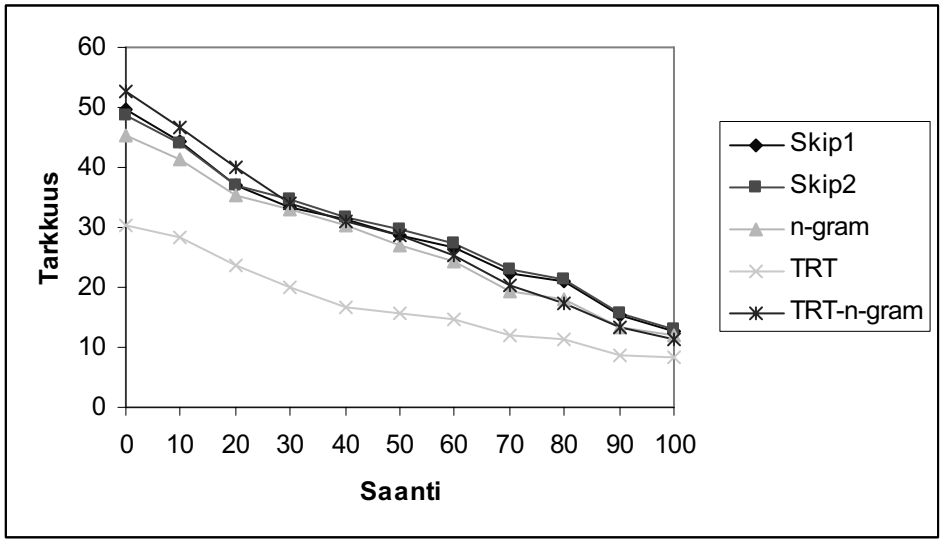
sumean käännösmenetelmän tunnistamiseksi. Kuten aiemmin todettiin, TRT-kysely suoriutui selvästi muita sumeita kyselytyyppejä heikommin. Ero TRT-kyselyn ja grammikyselyiden keskitarkkuuksissa oli useimmilla saantitasoilla tilastollisesti merkitsevä tai erittäin merkitsevä. Tämän vuoksi TRT-käännös ei yksinään vaikuta riittävältä kyselynkäännösmenetelmältä. Aiemmat tutkimustulokset (Toivonen ym. 2005) tukevat tätä johtopäätöstä, vaikka TRT-käännöksen suorituskyky olisi voinut osoittautua paremmaksi, jos käytössä ollut transformaatiosääntökokoelma olisi ollut onnistuneempi.

Tuloksissa eri grammimenetelmien välille ei syntynyt sellaisia eroja, jotka riittäisivät osoittamaan jonkin menetelmän selvästi muita paremmaksi: erot olivat prosentuaalisesti pieniä eivätkä tilastollisesti merkitseviä. Kuvan 1 saantitarkkuuskäyristä näkyy, kuinka yhdistetyn TRT-n-grammimenetelmän saavuttama keskitarkkuus oli matalilla saantitasoilla muita menetelmiä korkeampi. Saantitasoilla 30–60 kaikki grammimenetelmät olivat hyvin tasavertaisia. Korkeilla saantitasoilla s-grammit suoriutuivat muita sumeita menetelmiä paremmin, kuten saantitarkkuuskäyristä voidaan nähdä. Vaikka erot olivat pieniä ja eivätkä tilastollisesti merkitseviä, olivat s-grammit ja TRT-n-grammit keskimäärin hieman n-grammimenetelmää parempia. Transformaatiosääntöihin yhdistäminen parantaa n-grammien suorituskykyä hieman, vaikkakin odotettua vähemmän. Menetelmien yhdistäminen vaikuttaa siten jossain määrin hyödylliseltä lähestymistavalta läheisten sukulaiskielten välisessä tiedonhaussa. Menetelmien yhdistämien saattaisi saavuttaa parempia tuloksia, jos käytössä olisi laajemman aineiston pohjalta muodostettu transformaatiosääntökokoelma.

6. Johtopäätökset

Tämän tutkimuksen tavoitteena oli tutkia (1) ovatko sumeat käännösmenetelmät sanakirjakäännöksen kanssa kilpailukykyisiä käännösmenetelmiä läheisten sukulaiskielten välisessä tiedonhaussa, ja (2) mikä sumea käännösmenetelmä on lupaavin lähestymistapa läheisten sukulaiskielten välisessä tiedonhaussa. Viiden sumean merkkijonomenetelmän suorituskykyä testattiin kyselynkäännöksessä norjasta ruotsiin vuoden 2003 CLEF hakuaiheita käyttäen. Tutkimuksen tulokset olivat rohkaisevia ja antoivat tukea hypoteesille, jonka mukaan sanakirjaperustainen kyselynkäännösmenetelmä voitaisiin korvata yksinkertaisilla sumeilla merkkijonomenetelmillä lähisukulaiskielten välisessä tiedonhaussa.

Ero parhaiden sumeiden menetelmien ja sanakirjakäännöksen suorituskyvyn välillä ei ollut tilastollisesti merkitsevä. Paras sumea menetelmä (Skip2) saavutti keskimäärin 85 % sanakirjakäännöksen tehokkuudesta, mikä viittaa sumeiden käännösmenetelmien käyttökelpoisuuteen kyselynkäännöksessä lähikielten välillä. Kehittämällä menetelmiä edelleen niiden tehokkuutta voidaan parantaa. Tulokset eivät osoittaneet yhtä sumeaa menetelmää selkeästi parhaaksi käännösmenetelmäksi. Tilastollisiin käännössääntöihin perustuva TRT-käännös oli kuitenkin selvästi muita menetelmiä heikompi, eikä siten tässä tutkimuksessa osoittautunut yksinään riittäväksi käännösmenetelmäksi. Aiemmissa tutkimuksissa s-grammit on todettu n-grammeja tehokkaammaksi välineeksi kieltenvälisten kirjoitusasuvarianttien käsittelyssä (Pirkola ym. 2002; Keskustalo ym. 2003). Tässä tutkimuksessa menetelmien välille ei kuitenkaan syntynyt tilastollisesti merkitseviä eroja, vaikka s-grammit olivat keskimäärin n-



Kuvio 1. Saanti-tarkkuuskäyrät sumeilla menetelmillä käännetyille kyselytyypeille

grammeja parempia kaikilla saantitasoilla. Myös yhdistetty TRT-n-grammi -menetelmä antoi vastaavan tuloksen: TRT-n-grammit olivat keskimäärin n-grammeja parempia, mutta erot eivät olleet tilastollisesti merkitseviä. Ero n-grammien ja yhdistetyn menetelmän välillä oli pienempi kuin aikaisemman tutkimuksen perusteella odotettiin (ks. Toivonen ym. 2005).

Aiemmissa tutkimuksissa sumeita menetelmiä on käytetty nimenomaan sanakirjoista puuttuvan erityistermistön kääntämiseen (Pirkola ym. 2002; Keskustalo ym. 2003; Toivonen ym. 2005). Tässä tutkimuksessa sumeiden menetelmien käyttö laajennettiin kattamaan koko kyselynkäännös läheisten sukulaiskielten välillä, mikä on tehtävänä erityistermitön kääntämistä vaativampi. Erot aiempien tutkimusten tuloksiin saattavat johtua tästä. Toisaalta sumeiden käännösmenetelmien arviointi tässä tutkimuksessa eroaa aiemmista tutkimuksista, joissa menetelmiä on arvioitu pääasiassa niiden saavuttaman käännöstarkkuuden perusteella, ilman käännettyjen kyselyiden suorituskyvyn arviointia. Kaikkia tutkittuja grammimenetelmiä voidaan kuitenkin pitää kiinnostavina jatkotutkimuksen kannalta, s-grammien ja yhdistetyn TRT-n-grammimenetelmän ollessa vain hieman n-grammeja lupaavampia.

Pirkola ym. (2003) totesivat, että toimivan trans formaatiosääntökokoelman luominen vaatii laajan, tuhansien sanaparien vastinsanaparikokoelman käyttöä sääntöjen lähdeaineistona. Tässä tutkimuksessa vastinsanaparikokoelma jäi melko pieneksi (n. 3000 sanaparia) ja käytettävissä olevien transformaatiosääntöjen määrä siten alhaiseksi. Monille kyselyiden sanoille ei saatu tuotettua lainkaan käännösvastineita transformaatiosääntöjen avulla, jolloin kyselyissä käytettiin alkuperäisiä norjankielisiä hakusanoja. On lupaavaa, että TRT-kysely tästä huolimatta suoriutui 33,5 % kääntämätöntä norjankielistä kyselyä paremmin. Siten on mahdollista, että TRT-menetelmän suorituskyky paranisi onnistuneempaa transformaatiosääntökokoelmaa käytettäessä ja erityisesti yhdistetty TRT-n-grammi käännös osoittautuisi tehokkaaksi kysely nkäännösmenetelmäksi läheisten sukulaiskielten välillä. Tämän vuoksi sääntöjen luomiseen käytettävää vastinsanakokoelmaa tullaan jatkossa laajentamaan.

Myös transformaatiosääntöjen yhdistämistä s-grammeihin tullaan jatkossa tutkimaan, sillä yhdistelmän uskotaan parantavan sumean käännöksen toimivuutta edelleen. Toisaalta transformaatiosääntöjen grammikäännökseen tuoma lisäarvo riippuu tutkittavasta kieliparista: kielten ollessa hyvin samankaltaisia s-grammimenetelmä saattaa toimia hyvin ja riittää yksinään suurimman osan kirjoitusasuvarianteista tunnistamiseen. Tällöin transformaatiosäännöt eivät enää juuri paranna käännöksen tasoa (Pirkola ym. 2003). Siten myös kielten

ortografioiden samankaltaisuus saattaa selittää miksi transformaatiosäännöt eivät parantaneet n-grammi-käännöksen tulosta odotetulla tavalla norjan ja ruotsin välillä. Yhdistettyjä menetelmiä olisikin mielenkiintoista testata kielillä, joiden välillä kirjoitusasun variaatio on hieman näitä kieliä suurempaa. Tässä mielessä sumeiden käännösmenetelmien tutkimuksen laajentaminen koskemaan myös tanskan kieltä on kiinnostavaa: tanska ja ruotsi eroavat kirjoitusasultaan hieman norjaa ja ruotsia enemmän, kirjoitusasuvarianttien osuuden kielten sanastoista kuitenkin pysyessä korkeana. Samalla norja ja tanska ovat keskenään jopa norjaa ja ruotsia läheisempiä sukulaiskieliä, joiden välillä sumea käännös saattaisi toimia erittäin hyvin.

Tässä tutkimuksessa kaikkien kyselyiden sanat normalisoitiin, sillä transformaatiosäännöt eivät sisältäneet sääntöjä taivutusmuotojen käsittelyyn. Tämä ei kuitenkaan ole johdonmukainen lähestymistapa sanakirjariippumatt omuuteen pyrkivässä tutkimuksessa. Jatkossa luodaan sääntökokoelma, joka kykenee myös taivutusmuotojen käsittelyyn. Sanojen esiintymisfrekvenssiin perustuva FITE-TRT menetelmä (Pirkola ym. 2006) parantanee myös jatkossa transformaatiosääntöihin perustuvan käännöksen laatua. Sen avulla voidaan luotettavasti tunnistaa paras kohdekielinen käännösvastine lähdekieliselle sanalle.

Hyväksytty julkaistavaksi 27.9.2006.

Lähteet

Barðdal, J., Jörgensen, N., Larsen, G., & Martinussen B. 1997. Nordiska: Våra språk förr och nu. Lund, Studentlitteratur.

Broglio, J., Callan, J. & Croft B. 1993. Inquery system overwiev. Julkaisussa: TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, Virginia, September 19-23. s. 47-67. URL: http://acl.ldc.upenn.edu/X/X93/X93-1008.pdf (27.9.2006)

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. & Järvelin, K. 2004. Dictionary-based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000-2002. Information Retrieval 7(1-2): 99-119.

Keskustalo, H. & Pirkola, A. & Visala, K. & Leppänen, E., & Järvelin, K. 2003. Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants. Julkaisussa: Proceedings of the 10th International Symposium on String Processing and Information Retrieval, Manaus, Brazil, 8-10 October. Berlin, Springer, Lecture Notes in Computer Science 2857, 252 - 265.

Kishida, K. 2005. Technical issues of cross-language information retrieval: a review. Information Processing and Management 41(3): 433-455.

Kraaij, W. 2004. Variations on language modeling for information retrieval. PhD thesis, University of Twente.

McNamee, P. & Mayfield, J. 2003. JHU/APL Experiments in Tokenization and Non-Words Translation. CLEF 2003 Working Notes. URL: http://www.clef-campaign.org/2003/WN_web/03.pdf (27.9.06)

Nationalencyklopedin 1990, osa 14. Högnäs, Bokförlaget Bra Böcker.

O'Rourke, A., Robertson, A., & Willett, P. 1997. Word Variant Identification in Old French. Information Research 2 (4). URL: http://informationr.net/ir/2-4/paper22.html (27.9.06)

Peters, C. 2003. Introduction to the CLEF 2003 Working Notes. URL: http://www.clef-campaign.org/2003/WN_web/00.2%20-%20intro.pdf (27.9.06)

Pfeiffer, U., Poersch, T. & Fuhr, N. 1996. Retrieval effectiveness of proper name search methods. Information Processing & Management, 32(6): 667-679.

Pirkola, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. Julkaisussa: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, August 24-28. New York, ACM Press, 55-63.

Pirkola, A., Keskustalo H., Leppänen, E., Känsälä, A.P. & Järvelin, K. 2002. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. Information research, 7(2). URL: http://InformationR.net/ir/7-2/paper126.html (27.9.06)

Pirkola, A., Toivonen, J., Keskustalo, H. ,Visala, K. & Järvelin, K. 2003. Fuzzy Translation of Cross-Lingual Spelling Variants. Julkaisussa: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28– August 1. New York, ACM Press, 345 - 352.

Pirkola, A., Toivonen, J., Keskustalo, H. & Järvelin, K. 2006. FITE-TRT: A high quality translation

technique for OOV words. Julkaisussa: Proceedings of the 21st Annual ACM Symposium on Applied Computing. Dijon, France, April 23 -27. New York, ACM Press, s. 1043 - 1049. URL: http://www.info.uta.fi/tutkimus/fire/archive/2006/pirkola-FITE-TRT-SAC.pdf (27.9.06)

Robertson, A.M. & Willet, P. 1998. Applications of n-grams in textual information systems. Journal of Documentation, 54(1): 48-69.

Spark Jones, K. 1974. Automatic indexing. Journal of Documentation 30(4): 393-432.

Toivonen, J., Pirkola, A., Keskustalo, H., Visala, K., & Järvelin, K. 2005. Translating cross-lingual spelling variants using transformation rules. Information Processing & Management, 41(4): 859-872.

Liite: esimerkkejä käytetyistä kyselytyypeistä

Ruotsinkielinen vertailukohtakysely

#sum(christo packeterar tyska riksdagshus konstnär christo inslagning tyska riksdagshus)

Norjankielinen vertailukohtakysely

#sum(christo pakke tysk riksdagsbygning innpakking tysk riksdag berlin kunstner christo)

Sanakirjakysely

#sum(@christo #syn(paket packe bunt ask packa) #syn(tysk tyska) #syn(regerings stats stat statlig) dag #syn(byggnadsverk byggnad konstruktion hus) packning #syn(tysk tyska) #syn(regerings stats stat statlig) dag @berlin konstnär @christo);

s-gram-kysely (CCI {{0},{1,2}})

#sum(#syn(tyristor mchistori @christo @christos) #syn(paket packe @pakue @takke) #syn(tysk tysktysk @tyskl @otysk) #syn(riksdagsbyggnad riksdagsbevakning @riksdagsebatten @riksdagsrupp) #syn(inpackning inpassning @king @parking) #syn (tysk tysktysk @tyskl @otysk) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berglin merlin @berlin @berlins) #syn(konstnär konstcenter @kunstler @kunstlers) #syn(tyristor mchistori @christo @christos))

TRT kysely

#sum(#syn(christo) #syn(packa pakka packe pakke) #syn(tysk) #syn(riksdagsbygning) #syn(innpacking innpakking) #syn(tysk) #syn(riksdag) #syn(berlin) #syn(kunstner) #syn(christo))

Yhdistetty TRT-n-gram -kysely

#sum(#syn(mchistori chefshistorik @christo @christos) #syn(packa packad @packard @packalén) #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdagsbyggnad riksdagsbevakning @riksdagsoch @landsbygsriksdagen) #syn(inpackning inpacka @inpac @racking) #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berliner berlinsk @berlin @berlins) #syn(kungstiger kungakonst @kunst @kunstler) #syn(mchistori chefshistorik @christo @christos))