Tiina Rajala

Ruotsin kielen yhdyssanat ja niiden morfologinen käsittely tiedonhaussa

Informaatiotutkimuksen pro gradu -tutkielma
Huhtikuu 2008
Informaatiotutkimuksen laitos
Tampereen yliopisto

TAMPEREEN YLIOPISTO

Informaatiotutkimuksen laitos

RAJALA, TIINA: Ruotsin kielen yhdyssanat ja niiden morfologinen käsittely tiedonhaussa

Pro gradu -tutkielma, 104 s., 4 liites.

Informaatiotutkimus

Huhtikuu 2008

TIIVISTELMÄ

Tutkimuksen tarkoituksena on kartoittaa yhdyssanojen roolia ruotsin kielessä tiedonhaun näkökulmasta. Empiirisen tutkimuksen ensimmäisessä osassa selvitetään yhdyssanojen määrää ja tyyppejä ruotsinkielisissä hakuaiheissa ja dokumenttiotoksessa. Yhdyssanatyypeistä erityisen kiinnostuksen kohteena on kompositionaalisten ja ei-kompositionaalisten yhdyssanojen suhde. Empiirisen tutkimuksen toisessa osassa tarkoituksena on vertailla erilaisia kyselysarjoja ja selvittää, onko yhdyssanojen morfologisesta käsittelystä hyötyä ruotsinkielisessä tiedonhaussa. Tutkimus on perinteinen tiedonhaun laboratoriotutkimus, jossa verrataan kolmea eri kyselysarjaa sekä yhdyssanoiltaan eliminoidussa että yhdyssanoiltaan eliminoimattomassa hakemistossa. Kyselysarjat koostuvat kolmenlaisista kyselyistä: a) perusmuotoinen kysely, jonka yhdyssanat ovat osittamattomina, b) perusmuotoinen, yhdyssanoiltaan ositettu kysely, yhdysosia ei eliminoitu sekä c) perusmuotoinen, yhdyssanoiltaan ositettu ja yhdysosiltaan eliminoitu kysely. Kyselysarjoista on vertailussa sekä rakenteeton kysely että rakenteinen #syn- ja #uw20-operaattoreita hyödyntävä versio. Aineistoina käytetään vuosien 2002 (49 hakuaihetta) ja 2003 (54 hakuaihetta) CLEF-aineistoja sekä Per Ahlgrenin kokoelmaa (51 hakuaihetta). Kyselyjen ja hakemistojen morfologisessa käsittelyssä käytetään SWETWOL-analyysiohjelmaa. Tiedonhakujärjestelmänä toimii Indri.

Hakuaiheiden ja dokumenttiotoksen yhdyssanojen tutkiminen osoitti, että yhdyssanoilla on tärkeä rooli tiedonhaussa. Yhdyssanojen määrä todettiin sekä hakuaiheissa että dokumenteissa aiemmin esitettyjä arvioita suuremmaksi. Sanaluokkien osalta esiintyy eniten substantiivisia yhdyssanoja, jotka ovat myös tiedonhaussa merkityksellisiä hakuavaimia. Toiseksi eniten esiintyy yhdyssanaverbejä, joiden osiin jakaminen ei kuitenkaan ole hyödyllistä. Kompositionaaliset yhdyssanat saavuttivat sekä hakuaiheissa että dokumenteissa 60 prosentin enemmistön, mutta myöskään ei-kompositionaalisten yhdyssanojen roolia ei ole syytä unohtaa. Koska enemmistö kuitenkin oli kompositionaalisia yhdyssanoja, voitaisiin olettaa, että yhdyssanojen käsittely tiedonhaussa olisi ruotsin kielessä tarkoituksenmukaista.

Rakenteisten kyselysarjojen osalta tutkimus ei kuitenkaan antanut selkeää vastausta siihen, kannattaako yhdyssanoja käsitellä ruotsin kielessä kyselyvaiheessa. Erot menetelmien välillä olivat pääasiassa pieniä eivätkä tilastollisesti merkitseviä. CLEF2003-aineisto antoi viitteitä siitä, että yhdyssanoja kannattaisi käsitellä ruotsin kielessä tiedonhaussa. CLEF2002-aineisto antoi kuitenkin päinvastaisia tuloksia, ja tämän aineiston osalta perusmuotoinen, yhdyssanoiltaan osittamaton kyselysarja oli parhaiten menestyvä menetelmä. Per Ahlgrenin kokoelman kyselysarjojen osalta yhdyssanojen osittaminen ilman yhdysosien eliminointia huononsi tarkkuusarvoja lähes kaikilla relevanssitasoilla. Yhdyssanojen osittaminen eliminointia hyödyntäen paransi tark-

kuusarvoja kaikilla relevanssitasoilla, mutta erot perusmuotoiseen, osittamattomaan kyselysarjaan olivat hyvin pieniä. Rakenteisten kyselysarjojen osalta ei ole täysin selvää, missä määrin läheisyysoperaattoreiden käyttö vaikuttaa hakutuloksiin.

Hakuaihetasolla analysoitaessa rakenteisissa kyselyissä oli havaittavissa hakuaihekohtaista vaihtelua siinä, kannattaako yhdyssanoja käsitellä ruotsinkielisessä tiedonhaussa. Kaiken kaikkiaan hakuaihetason analyysin perusteella vaikuttaa siltä, että yhdyssanojen luonne ratkaisee, onko niiden käsitteleminen hyödyllistä vai ei. Kaikkia kompositionaalisiakaan yhdyssanoja ei ole hyödyllistä osittaa. Paikoitellen SWETWOL:n tekemillä virhetulkinnoilla oli myös vaikutusta eri kyselysarjojen menestymiseen.

Rakenteettomien kyselysarjojen tulokset puolestaan antoivat viitteitä siitä, että varsinkin yhdyssanojen osittaminen eliminointiperiaatetta hyödyntäen on tehokas menetelmä ruotsin kielessä. Kaiken kaikkiaan rakenteettomilla kyselysarjoilla saadut tarkkuusarvot olivat selvästi rakenteisilla kyselysarjoilla saatuja arvoja korkeampia, ja tilastollisesti merkitseviä eroja havaittiin erityisesti aineistojen rakenteisten ja rakenteettomien eliminoitujen kyselyiden välillä. Rakenteettomien kyselysarjojen osalta kaikissa aineistoissa eliminointi oli tarkkuusarvoiltaan paras menetelmä. Varsinkin CLEF2003- ja Ahlgren-kyselysarjoilla erot olivat kohtuullisen suuria, mutta eivät tilastollisesti merkitseviä. Myös yhdyssanojen osittaminen ilman eliminointia menestyi hyvin CLEF-aineistoissa, kun taas Ahlgren-kyselysarjojen osalta menetelmä menestyi huonoiten lähes kaikilla relevanssitasoilla.

Myös rakenteettomien kyselyiden hakuaihekohtainen vertailu vahvisti sitä, että yhdyssanoiltaan ositetut ja eliminoidut kyselyt paransivat eri hakuaiheiden tarkkuusarvoja useammin kuin eliminoimattomat kyselyt. Rakenteisten kyselyjen tavoin myös rakenteettomat kyselyt tarjosivat esimerkkejä kompositionaalisista yhdyssanoista, joita ei ole kannattavaa osittaa. Varsinkin eliminoimattomissa kyselyissä SWETWOL:n tekemät virhetulkinnat huononsivat tarkkuusarvoja.

Sisällys

1 JOHDANTO

Luonnollinen kieli on olennainen osa ihmisten välistä vuorovaikutusta. Kielellä on myös suuri painoarvo tietoa hankittaessa ja haettaessa. Tiedonhakija ilmaisee tiedontarpeensa luonnollista kieltä käyttäen. Toisaalta hakujärjestelmät toimivat vain merkkijonojen tasolla. Niille tiedonhaun onnistumisen edellytys on merkkijonojen välinen samankaltaisuus. Luonnollisen kielen taso on näin ollen niille vieras. Jotta hakujärjestelmät osaisivat tulkita tiedonhakijan tiedontarpeita, on ne esitettävä merkkijonoina kyselyn muodossa. Tiedonhakuun vaikuttaa myös se, että luonnollinen kieli on vaihtelevaa ja rikasta. Kaikkea kieleen liittyvää vaihtelua on mahdotonta ottaa huomioon tiedonhaussa. Kielen vivahteikkuuden vuoksi samasta asiasta käytetään erilaisia ilmaisuja, joten tiedonhakijan ja dokumentin kirjoittajan käyttämät sanat eivät merkkijonoina välttämättä täsmää toisiinsa. Tiedonhakuja voidaan myös tehdä eri kielillä, ja kieletkin eroavat toisistaan monin tavoin. Alkula (2000: 48) toteaakin väitöskirjassaan tiedonhaun yhdeksi ongelmakohdaksi sen, ”miten merkitykseltään lähekkäiset, mutta merkkijonoina erilaiset sanat ja sananmuodot saataisiin yhdistettyä (conflation) toisiinsa”.

Kun puhutaan luonnollisesta kielestä tiedonhaun tutkimuskentässä, englannin kieli on pitkään ollut tutkimuksen pääkohteena, koska englanti on tiedonhaun valtakieli. Kielet eivät kuitenkaan ole kaikin puolin samanlaisia, vaan eri kieliin liittyy erilaisia kielellisiä ilmiöitä, jotka aiheuttavat haasteita tiedonhaussa. Siinä missä esimerkiksi englannin kielessä sanaliittojen, eli erikseen kirjoitettujen sanakokonaisuuksien (information retrieval), tunnistaminen on tiedonhaun ongelmakohta, ruotsin kielessä suurempia haasteita aiheuttavat yhdyssanat ja erityisesti niiden loppuosien tunnistaminen hakuavaimina. 2000-luvulla kiinnostus kielellisiin ilmiöihin myös muiden kielten kuin englannin osalta onkin lisääntynyt. Esimerkiksi ruotsin, suomen ja saksan kielten osalta on tehty kielellisiä ilmiöitä koskevia tiedonhaun tutkimuksia (Alkula 2000; Braschler & Ripplinger 2004; Hedlund et al. 2001; Hedlund 2003).

Tutkimukseni kohteena on ruotsin kieli. Germaanisten kielten alahaaraan kuuluvaa kieltä puhuu äidinkielenään 8,5 miljoonaa ihmistä (Karlsson 2006: 264). Lisäksi ruotsin kieltä käytetään paljon Pohjoismaissa viestinnän kielenä. Ruotsin kieleen liittyy monia tiedonhaun näkökulmasta haasteellisia ilmiöitä. Yksi tällainen ilmiö on se, että ruotsin kielessä esiintyy runsaasti yhdyssanoja. Yhdyssanojen loppuosa sisältää usein sanan päämerkityksen, mutta jää haussa piiloon. Siksi on tarpeen selvittää, millainen yhdyssanojen rooli on ruotsinkielisessä tiedonhaussa ja millai-

sia menetelmiä yhdyssanojen käsittelemiseen on olemassa. Tutkielmassa tarkoituksena on tutkia yhdyssanoja varten kehitettyä kielenkäsittelymenetelmää eli sitä, kannattaako yhdyssanat jakaa osiin kyselyissä ruotsin kieltä koskevassa tiedonhaussa. Menetelmän tehokkuutta tutkitaan vertailemalla erilaisia kyselynmuodostustapoja. Lisäksi pyritään kartoittamaan yhdyssanojen esiintymiä ja tyyppejä ruotsinkielisissä hakuaiheissa ja dokumenteissa.

2 TIEDONHAUN KÄSITTEISTÖÄ

Tutkielmassa tehtävä tutkimus sijoittuu tiedonhaun tutkimusalaan. Informaatiotieteissä tehdään selkeä ero tiedonhankinnan ja tiedonhaun välille. Tiedonhankintatutkimus on kiinnostunut muun muassa ihmisten tiedontarpeista, heidän käyttämistään tiedonhankintakanavista sekä heidän saamansa tiedon hyödyllisyydestä. Tiedonhausta puhuttaessa liikutaankin mikrotasolla ja ollaan kiinnostuneita tiedonhakijan ja hakujärjestelmän välisestä vuorovaikutuksesta. Tiedonhaun tutkimus kattaa kuitenkin tiedonhaun lisäksi myös tiedon tallennuksen, organisoinnin ja järjestämisen tutkimuksen. (Alaterä & Halttunen 2002: 13−14.) Ingwersenin mukaan tiedonhaun tutkimuksessa kiinnostuksen kohteena ovat tiedonhakijalle hyödyllisen tiedon esittämiseen, tallentamiseen, hakemiseen ja löytämiseen liittyvät prosessit. Tiedonhaun tutkimuksen tarkoituksena on tutkia tiedonhaun prosesseja, jotta voitaisiin kehittää ja testata hakujärjestelmiä, jotka mahdollistavat onnistuneen halutun tiedon kommunikoinnin tiedontuottajan ja tiedonkäyttäjän välillä. (Ingwersen 1992: 49.) Tämän tutkielman tutkimus on puolestaan tiedonhaun laboratoriotutkimus, jossa käyttäjä tai tiedonhakija ei ole mukana, vaan jossa hakuaiheet edustavat käyttäjiä. Tiedonhaun laboratoriotutkimusta esitellään tarkemmin tämän luvun alaluvussa 2.6. Tämän lisäksi luvussa esitellään tiedonhaun näkökulmasta keskeisiä käsitteitä ja ilmiöitä.

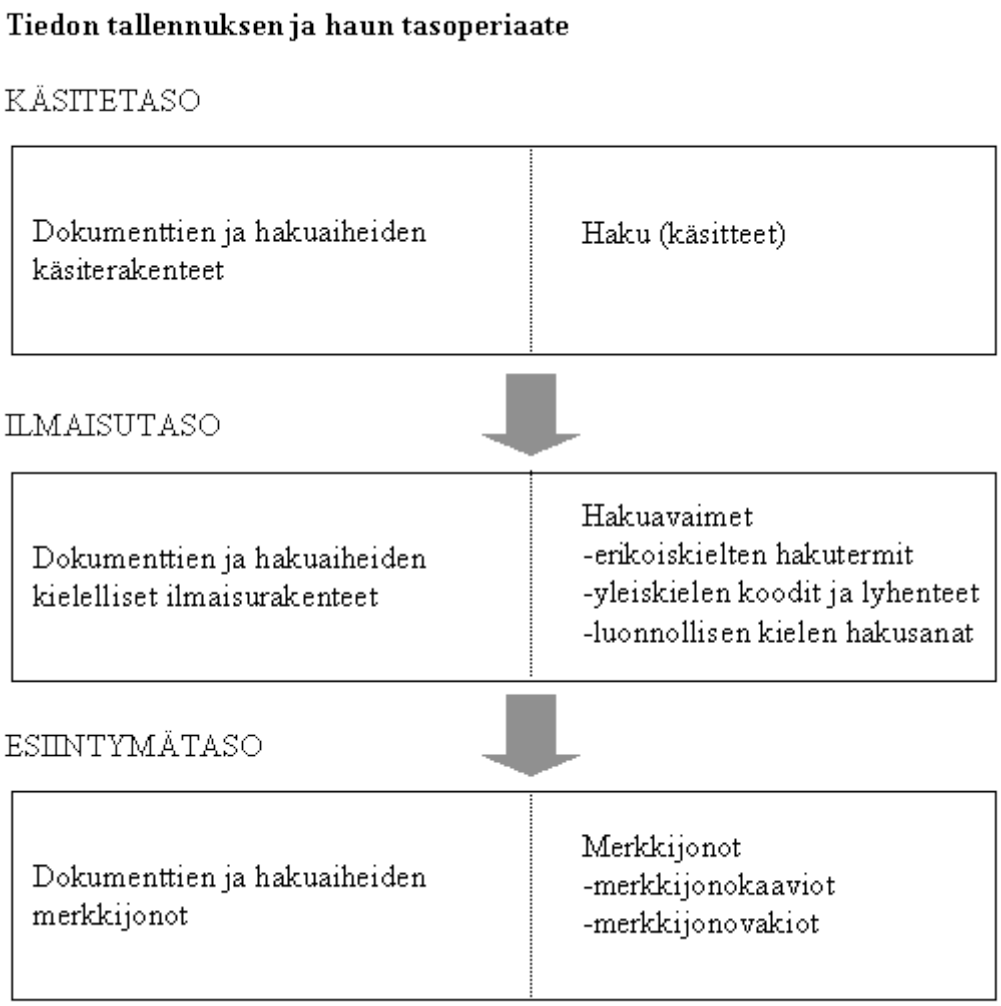
2.1 Tiedonhaun lähestymistavat

Tiedonhakua voidaan lähestyä useista eri näkökulmista. Järvelin (1995: 30−33) esittelee neljä eri näkökulmaa. *Tiedonhaku täsmäyttämisenä* keskittyy hakuaiheiden ja dokumenttien esitysten täsmäyttämiseen täsmäytysmekanismia apua käyttäen. *Tekninen prosessinäkökulma* huomioi tiedonhaun konkreettiset vaiheet ja välineet, eli millaisia vaiheita tiedonhakuprosessiin kuuluu ja mitä välineitä prosessi vaatii. *Kognitiivisen näkökulman* kiinnostuksen kohteena on ihminen, tiedontarvitsija, tietämysrakenteineen. Kiinnostavaa tästä näkökulmasta ovat erityisesti tiedonha-

kuun liittyvä ajatustyö, tiedonkäsittely ja henkilön tietämysrakenteissa tapahtuvat muutokset. *Evaluoiva näkökulma* puolestaan keskittyy tiedonhaun tuloksellisuuteen ja kustannustehokkuuteen. Tiedonhakuprosessia voidaan tarkastella joko makro- tai mikrotasolla. Makrotasolla huomion kohteina ovat tiedonhaun kokonaisprosessin tuottamat tulokset, mikrotasolla puolestaan keskitytään hakuprosessiin vaikuttaviin eri tekijöihin sekä kokonaisprosessin eri vaiheiden tuloksellisuuteen. (Järvelin 1995: 30−33.) Tässä tutkielmassa tehtävä tutkimus onkin näkökulmaltaan evaluoiva, sillä tutkimuksessa pyritään selvittämään kyselynmuodostustapojen tehokkuutta.

2.2 Tiedon tallennuksen ja haun tasoperiaate

Järvelin (1995: 68) tiivistää teoksessaan näin: ”Informaatio on käsiterakenne, joka ilmaistaan kielen avulla dokumenttina, joka tallennetaan datana hakujärjestelmään.” *Tiedon tallennuksen ja haun tasoperiaatteen* ideana onkin se, että niin tiedontallennuksessa kuin tiedonhaussakin on osallisena kolme eri tasoa. Näiden kolmen eri tason olemassaolon ymmärtäminen on tärkeää, jotta tiedonhaku on tuloksellista. Sekä dokumentteja että tiedontarpeita voidaan lähestyä kolmesta eri näkökulmasta. Teknisestä näkökulmasta dokumentit sisältävät merkkijonoja (esiintymätaso), jotka puolestaan esittävät luonnollisen kielen ilmaisuja (ilmaisutaso), ja nämä puolestaan edustavat dokumentin käsitteellistä sisältöä (käsitetaso). (Järvelin & Sormunen 1999: 124.) Kuvassa 1 on esitetty tiedon tallennuksen ja haun tasoperiaate. Vasemmalla puolella on selitetty kukin taso esimerkein, oikealla puolella puolestaan esitetään, mitä tasot tarkoittavat käytännön tiedonhaussa.



KUVA 1. Tiedon tallennuksen ja haun tasoperiaate Järveliniä (1995: 69, 177) mukaillen.

2.2.1 Käsitetaso

Käsitetasolla liikutaan kognitiivisella tasolla, ja käsitetasolla tarkoitetaankin tiedonhakijan tai tiedontuottajan käsiterakenteita. Käsiterakenteet ovat mukana sekä tietoa haettaessa että tallennettaessa. Niillä tarkoitetaan hakuaiheen tai dokumentin käsitteitä ja niiden välisiä suhteita. Kummassakin vaiheessa tapahtuu joko dokumentin tai hakuaiheen käsiteanalyysi, jolloin selvitetään aiheeseen liittyvät käsitteet ja niiden väliset suhteet. Esimerkiksi tiedonhaussa käsitetasossa onkin kysymys tiedonhakijan tiedontarpeesta ja hänen aiheeseen liittyvistä käsiterakenteistaan. (Järvelin 1995: 69)

Dokumentin tuottajalla on taustalla joku käsitteellinen sisältö, jonka hän haluaa ilmaista lukijalle. Lukijalle tämä sisältö paljastuu luonnollisen kielen muodossa, jonka ilmaisut puolestaan esiintyvät hakujärjestelmässä merkkijonoina. Sekä dokumentin kirjoittajalla että lukijalla on omat nä-

kökantansa lähestyä dokumentin aihetta. Tiedonhakijan onkin otettava huomioon eri kirjoittajien mahdolliset erilaiset käsitteelliset näkökulmat, erilaiset ilmaisut käytetyille käsitteille sekä tietokoneen erilaiset tavat käsitellä kyseisiä ilmaisuja merkkijonoina. (Järvelin & Sormunen 1999: 124−126.)

2.2.2 Ilmaisutaso

Ilmaisutaso eli lingvistinen taso puolestaan viittaa siihen, miten käsitteet ilmaistaan luonnollisella kielellä tai jollain erikoiskielellä, esimerkiksi dokumentaatiokielellä. Käytännössä tämä tarkoittaa siis hakuaiheen käsitteiden esityksiä ilmaisutasolla eli haussa käytettäviä hakuavaimia. (Järvelin 1995: 70−71) Vastaavasti tiedontallennusvaiheessa tiedontallentajan käsiterakenteet muokkautuvat indeksoinnissa luonnolliseksi kieleksi tai dokumentaatiokieleksi, jonka avulla dokumentin sisältöä kuvataan.

Kuvassa 1 on esitetty ilmaisutaso tiedonhaun näkökulmasta. *Hakusanoilla* tarkoitetaan luonnollisen kielen hakuilmaisuja eli yksittäisiä sanoja tai yhdyssanoja. *Hakutermi* puolestaan on erikoiskielten termi, joka voi olla sanaperusteinen tai koodiperusteinen. Sanaperusteisen hakutermin sanat ovat luonnollista kieltä, kun taas koodiperusteinen hakutermi ei kuulu luonnollisen kielen ilmaisuihin (esimerkiksi kun haku kohdistetaan luokkaan). *Hakuavain* on yhteisnimitys, jota käytetään hakusanoista, hakutermeistä sekä yleiskieleen sisältyvistä lyhenteistä ja koodeista. (Järvelin 1995: 176−177.) Tässä tutkielmassa käytetäänkin yhteisnimitystä hakuavain.

2.2.3 Esiintymätaso

Kolmas taso on *esiintymätaso*, jolla viitataan merkkijonoihin. Esiintymätaso on keskeinen osa tiedonhakua, koska hakujärjestelmät käsittelevät ainoastaan merkkijonoja. Hakujärjestelmät keskittyvät ainoastaan merkkijonojen keskinäiseen samanlaisuuteen, niiden esiintymien keskinäiseen sijaintiin (etäisyys), sijaintirakenteeseen (missä kentässä esiintyy) sekä esiintymien lukumäärään. Tiedontallennusvaiheessa dokumentti esitetään datana ja tallennetaan merkkijonoina tietokantaan. Tiedonhaussa hakuaihe esitetään kyselynä, joka rakentuu merkkijonoista. Kyselyä muotoiltaessa ilmaisutason hakuavaimet siis käännetään esiintymätason merkkijonoiksi. (Järvelin 1995: 72.)

Tiedonhaun näkökulmasta merkittävää on myös se, että kun käsitetason käsiterakenteet ja ilmaisutason hakuavaimet esitetään merkkijonoina, ne irrotetaan tekstiyhteydestään eli kontekstistaan (Järvelin 1995: 72). Tällöin esimerkiksi luonnolliseen kieleen liittyvät ongelmat tulevat näkyviksi. Hakujärjestelmät eivät ymmärrä merkkijonoista muodostuvien kokonaisuuksien merkityksiä. Tiedonhakijan on osattava hakea täsmällisillä merkkijonoilla ja huomioitava esimerkiksi sanojen taipuminen ja muut kieleen liittyvät ongelmakohdat, joita käsitellään tarkemmin seuraavassa luvussa.

Kuvassa 1 on esitetty esiintymätaso tiedonhaun näkökulmasta. Esiintymätason merkkijonot ovat nimensä mukaisesti merkeistä koostuvia jonoja. *Merkkijonovakiot* ovat ilmaisutason hakuavaimien erilaisia muotoja. Sanat taipuvat ja samasta sanasta esiintyy hakujärjestelmän hakemistossa erilaisia taivutusvariantteja, joita merkkijonovakiot esiintymätasolla siis edustavat. Sanoista on myös erilaisia kirjoitusasuja ja sananmuotoja johdosten ja yhdyssanojen vuoksi. *Merkkijonokaavioita* käytetäänkin tämän vuoksi, koska ne täsmäävät useisiin eri merkkijonovakioihin. Esimerkki merkkijonokaaviosta on *merkkijonon katkaisu* (truncation). (Järvelin 1995: 178, 198−199)

Katkaisemiseen käytetään *jokerimerkkiä*. Katkaisemista voidaan käyttää merkkijonon edessä. Esimerkiksi merkkijonolla **#industri** löydetään muun muassa sanat **skogsindustri** ja **vatteninindustri**. Yleisempi tapa on katkaista merkkijono lopusta (**rederi#**), jolloin löydetään hakuavaimen kaikki erilaiset taivutus- ja yhdyssanavariantit. Kaikkia kirjoitusasuja ei kuitenkaan löydy katkaisemalla merkkijono lopusta. Esimerkiksi suomen kielessä on runsaasti taipuvia sanoja (**yö**, **öitä**), jotka eivät löydy katkaisun avulla. Katkaisemisen lisäksi voidaan käyttää *merkin korvausta* (masking). Esimerkiksi korvaamalla yksi merkki sanasta **fader** (**f#der**) löydetään sekä perusmuoto **fader** että taivutusmuoto **fäder**. (Alaterä & Halttunen 2002: 43.) *Merkkijonon korvausta* (string masking) voidaan käyttää eripituisten etu- ja loppuliitteiden korvaamiseen. Lisäksi merkkijonon korvauksesta voi olla apua yhdyssanojen osia haettaessa sekä silloin, kun hakuavaimen kirjoitusasut ovat erilaisia merkkien määrän suhteen. Esimerkiksi **alumin?m** täsmää sekä **aluminium** että **aluminum** merkkijonoihin. (Järvelin 1995: 199−200)

Läheisyysoperaatioilla (adjacency operations, proximity operations) tai *läheisyysoperaattoreilla* (adjacency operator, proximity operator) voidaan tarkentaa hakuehtoja määrittelemällä esimerkiksi hakuavainten sijaintia ja etäisyyttä toisiinsa verrattuna (Alaterä & Halttunen 2002: 43). Läheisyysoperaattoreiden käyttö edellyttää sitä, että hakujärjestelmän hakemisto on jo tallennus-

vaiheessa laadittu siten, että käänteistiedostoon (ks. luku 2.3) on tietuenumeroiden lisäksi tallennettu myös tieto merkkijonojen sijainnista tietueessa (Järvelin 1995: 202).

Läheisyysoperaattoreiden avulla voidaan määritellä, miten hakuavaimet esiintyvät toisiinsa nähden. Ei riitä, että ne esiintyvät samassa dokumentissa. Läheisyysoperaattoreiden avulla hakuavainten välinen etäisyys voidaan määritellä sellaiseksi, että ne todennäköisesti kuuluvat vielä samaan tekstiyhteyteen eli kontekstiin. Läheisyysoperaattorit mahdollistavatkin sen, että satunnaiset virheelliset kytkennät hakuavainten välillä vältetään. Erilaisia läheisyysoperaattoreita ovat Järvelinin mukaan muun muassa seuraavat:

- merkkijonojen esiintyminen samassa kappaleessa
- merkkijonojen esiintyminen samassa lauseessa
- merkkijonojen esiintyminen samassa kentässä
- merkkijonojen esiintyminen n:n sanan etäisyydellä
- merkkijonojen esiintyminen vierekkäin

Eri hakujärjestelmät eroavat toisistaan siinä, millaisia läheisyysoperaattoreita ne mahdollistavat. (Järvelin 1995: 201−203.)

2.3 Tietokannan rakenne

Tiedonhaussa on kyse tiedonhakijan ja tiedonhakujärjestelmän välisestä vuorovaikutuksesta. Tiedonhaun toinen keskeinen osapuoli on siis *tiedonhakujärjestelmä*, jota käytetään tietoyksiköiden tallentamiseen, etsintään, jälleenhakuun ja jakeluun (Järvelin 1995: 20). Tiedonhakujärjestelmään liittyy *tietokanta*, joka on ”kokoelma tiettyä kohdetta kuvaavia tietoja, joita yksi tai useampi tietojärjestelmä käyttää ja päivittää” (MOT Atk-sanakirja 1.0 2007). Tietokannan perusyksikkö on *tietue* (record), joka sisältää yhden kuvailtavan kohteen tiedot. Tietue jakaantuu *kenttiin* ja *osakenttiin*, jotka kukin kuvaavat jotakin osaa kohteesta. Tällaisia kenttiä ovat esimerkiksi tekijä, nimeke ja indeksointitiedot. Tekstitietokannoissa tietueita voidaan kutsua myös *dokumenteiksi*. Tietueiden ja kenttien lisäksi tietokanta koostuu *tiedostoista*, jotka koostuvat useista tietueista. (Alaterä & Halttunen 2002: 16.)

Tiedonhaun nopeuttamiseksi useimmissa tiedonhakujärjestelmissä hyödynnetään *käänteisrakennetta* (inverted file structure), jossa tietueiden eri kentissä esiintyvät hakuavaimet esitetään aakkosjärjestyksessä olevana listana. Niiden yhteydessä ilmaistaan myös osoitteet kaikkiin niihin tietueiden numeroihin, joissa hakuavaimet esiintyvät. (Järvelin 1995: 96−97.) Näin muodostunut

käänteistiedosto (hakemisto, basic index, inverted file) sisältää siis kaikki tietueissa esiintyneet hakuavaimet sekä osoitteet tietueisiin. Käänteistiedostoa kutsutaan usein myös nimillä *hakemisto* tai *indeksi*. Vain haettavissa olevien kenttien sisältö indeksoidaan hakemistoon. (Alaterä & Halttunen 2002: 31.)

Baeza-Yates ja Ribeiro-Neto (1999: 192) kutsuvat hakuavainlistan yhteydessä esiintyviä tietuenumeroviittauksia esiintymiksi (occurrences). *Esiintymät* voivat sisältää lisäksi tiedon sekä merkkijonojen sijainneista (word positions) että merkkien sijainneista (character positions). Merkkijonojen ja merkkien sijainnin paikantaminen helpottaa fraasi- ja läheisyysoperaattoreiden käyttöä kyselyissä, koska sijaintilista ilmaisee tarkasti, missä kentässä ja kentän osassa hakuavain esiintyy. (Baeza-Yates & Ribeiro-Neto 1999: 192−193.)

Käänteishakemistoja muodostettaessa huomioidaan, mitä kenttiä asetetaan haettavaksi, mitä merkkijonoja otetaan mukaan (poistetaan sulkusanat eli stop words) ja kuinka kentät indeksoidaan. Sulkusanoilla tarkoitetaan tiedonhaun kannalta merkityksettömiä merkkijonoja, jotka esiintyvät dokumenteissa yleisesti. Sanaindeksoinnissa indeksointi tapahtuu sana sanalta, fraasiindeksoinnissa indeksoidaan fraasit eli sanaliitot, jolloin niiden sisältämät yksittäiset sanat eivät löydy yksinään. On myös mahdollista käyttää yhdistettyä sana- ja fraasi-indeksointia, jolloin hakuja voidaan tehdä sekä yksittäisillä sanoilla että sanaliitoilla. (Järvelin 1995:208.)

Kaikki nämä valinnat vaikuttavat tiedonhaun onnistumiseen ja siihen, missä muodossa kyselyt kannattaa esittää. Hakemistoja muodostettaessa on mahdollista hyödyntää erilaisia kielenkäsittelytapoja, joita esitellään tarkemmin luvussa 3.5. Mikäli hakemiston sanoja ei käsitellä mitenkään, ne esiintyvät hakemistossa taivutusmuotoisina (taivutusmuotoinen hakemisto), jolloin tiedonhaussa on hyödyllistä käyttää katkaisua. Sananmuotojen käsittelyllä voidaan saada aikaan esimerkiksi perusmuotoinen hakemisto, jonka sisältämät sanat ovat perusmuotoisia. Tällöin tiedonhaussakin voidaan käyttää perusmuodossa olevia hakuavaimia.

2.4 Täsmäytysmenetelmät

Tiedonhakujärjestelmässä tärkeässä roolissa on *täsmäytysalgoritmi*, jonka avulla lasketaan kyselyn ja dokumentin esityksen samankaltaisuus. Käytännössä tämä tarkoittaa sitä, kuinka hyvin kyselyn hakuavaimet täsmäävät dokumenttien indeksoinnissa käytettyihin hakemistosanoihin. Samankaltaisuuden perusteella ratkaistaan, kuuluuko dokumentti tulosjoukkoon ja mihin kohtaan

tulosjoukkoa dokumentti sijoittuu. *Täsmäytysmenetelmä* vertaa dokumenttien esityksiä kyselyn esitykseen. (Järvelin & Sormunen 1999: 122.) Tiedonhaun täsmäytysmenetelmät jakautuvat *täydellisen täsmäytyksen* (exact match) ja *osittaistäsmäytyksen* (partial match) menetelmiin.

Boolen logiikkaan perustuva hakumenetelmä kuuluu täydellisen täsmäytyksen menetelmiin. Boolen haussa käytetään ja-, tai- ja ei-operaattoreita, joiden avulla muodostetaan kyselylausekkeita. Hakutulos jakautuu kahtia niihin dokumentteihin, jotka täsmäävät kyselyyn, ja niihin dokumentteihin, jotka eivät täsmää. Hakumenetelmän ongelmana onkin Järvelinin mukaan se, että hakuun osittain tai melkein täsmääviä dokumentteja ei löydetä. Toinen ongelmakohta liittyy hakutuloksen esittämiseen, sillä Boolen haussa tulokseksi saadut dokumentit eivät järjesty relevanssin mukaiseen järjestykseen vaan täysin satunnaisesti. Näiden lisäksi Boolen operaattoreiden hallitseminen voi olla tiedonhakijalle hankalaa ja lisätä virheitä kyselyn muotoilussa. (Järvelin 1995: 107−108)

Täydellisen täsmäytyksen menetelmien ongelmien ratkaiseminen on vaikeaa ja tämän vuoksi tiedonhaun tutkimuksessa on suuntauduttu enemmän osittaistäsmäytystä hyödyntäviin menetelmiin. Näitä ovat muun muassa vektorimalliin (vector space model), sumeisiin joukkoihin (fuzzy set), todennäköisyyslaskelmiin (probability-based) ja nimikirjoitustiedostoihin (signature file) perustuvat menetelmät. (Järvelin 1995: 108) Osittaistäsmäytyksessä hakutuloksen kannalta olennaista on se, mitkä sanat kuvaavat parhaiten dokumentteja. Tausta-ajatuksen mukaan tällainen sana on yleinen dokumentissa, mutta harvinainen tietokannassa. Sanoille annetaan indeksoitaessa dokumenttikohtaiset painot, jotka määrittelevät niiden tilastollista edustavuutta dokumenttien kuvaamisessa. Kyselyn ja dokumentin yhteisten sanojen painojen avulla dokumentille lasketaan vertailuluku, ja näiden vertailulukujen perusteella hakutuloksen dokumentit järjestetään relevanttiuden mukaiseen järjestykseen. (Järvelin & Sormunen 1999: 122) Tässä tutkimuksessa käytetään osittaistäsmäytysmenetelmänä kielimallia (language modelling).

2.5 Tiedonhaun evaluointi

Tiedonhaun evaluoinnissa arvioinnin kohteina voivat olla muun muassa hakujärjestelmän toiminnallisuus, hakuun käytetty aika ja hakujärjestelmän komponenttien vaatima tila. Hakujärjestelmän suorituskyvyn arvioinnissa kiinnitetään huomiota muun muassa käytettävissä oleviin indeksointistrategioihin sekä hakujärjestelmän käyttäytymiseen ja sen aiheuttamiin kustannuksiin. Haun suorituskyvyn arvioinnissa puolestaan ollaan kiinnostuneita relevanttien dokumenttien

sijoittumisesta tuloslistalle ja haun tarkkuudesta. (Baeza-Yates & Ribeiro-Neto 1999: 73−74) Tässä alaluvussa keskitytään juuri haun suorituskyvyn arviointiin ja siihen, millaisia erilaisia mittareita suorituskyvyn arviointiin on olemassa.

2.5.1 Relevanssi

Tiedonhaun tulosten arvioimisessa peruskäsite on *relevanssi*. Relevanssi on käsitteenä monitulkintainen, mutta yleisesti sillä tarkoitetaan löydetyn tiedon hyödyllisyyttä. Kuten tiedonhakua myös relevanssia voidaan lähestyä useista eri näkökulmista. *Algoritminen relevanssi* kuvaa hakujärjestelmän algoritmin kykyä löytää kyselylle relevantteja dokumentteja. Tärkeintä on siis se, miten hyvin hakuavaimet täsmäävät dokumenteissa esiintyviin ilmaisuihin. *Aiherelevanssi* vastaa kyselyssä esitetyn aiheen ja dokumentin käsittelemän aiheen välistä suhdetta. Relevanssi toteutuu, jos löydetty dokumentti käsittelee kyselyssä esitettyä aihetta (aboutness). (Saracevic 1996: 214.)

Tiedonhaussa on esitetty vaihtoehtoja aiherelevanssille huomioimalla itse aiheen lisäksi myös käyttäjän tietämyksen tila sekä tilanne, josta tiedontarve saa alkunsa. Esimerkiksi sinänsä aihetta käsittelevä dokumentti ei välttämättä ole käyttäjälle hyödyllinen, jos hän on jo tutustunut siihen aikaisemmin tai jos dokumentti käsittelee aihetta väärästä näkökulmasta tai väärällä kielellä. Tämä kahtiajako (aihe vs. käyttäjä) onkin herättänyt paljon keskustelua eri tutkimussuuntauksissa. *Kognitiivinen relevanssi* tarkoittaakin tiedonhakijan tietämyksen tilan ja tiedontarpeen sekä löydettyjen dokumenttien välistä suhdetta. Relevanttien dokumenttien kriteereinä ovat dokumenteissa esitetyn tiedon kognitiivinen vastaavuus, informatiivisuus, uutuus ja laatu. *Tilannerelevanssi* puolestaan vastaa tiedonhakuun johtavan ongelman, tilanteen tai tehtävän sekä löydettyjen dokumenttien välistä suhdetta. Tärkeitä kriteereitä relevanssin arvioimisessa ovat saadun tiedon hyödyllisyys ja soveltuvuus ongelmatilanteen ratkaisemiseen ja epävarmuuden poistamiseen. *Affektiivinen relevanssi* vastaa tiedonhakijan aikomusten, tavoitteiden ja motivaation sekä löydettyjen dokumenttien välistä suhdetta. Tiedon arviointikriteereinä toimivat tiedonhakijan tyytyväisyys ja menestyminen. (Saracevic 1996: 214.)

Relevanssin monitulkintaisesta luonteesta huolimatta monissa tiedonhaun tutkimuksissa relevanssia lähestytään aiherelevanssin näkökulmasta. Dokumentit arvioidaan binäärisesti joko aiheeltaan relevanteiksi tai epärelevanteiksi. Relevanttien dokumenttien välillä voi siis olla suuria eroja siinä, kuinka laajasti ne käsittelevät kyselyn esittämää aihetta. Relevanssia voidaan kuiten-

kin arvioida myös moniportaisesti. Moniportaisissa relevanssiarvioissa huomioidaan myös se, kuinka perusteellisesti dokumentit käsittelevät aihetta. Dokumentit jaetaankin epärelevantteihin, marginaalisesti relevantteihin, jonkin verran relevantteihin sekä erittäin relevantteihin. Ahlgrenin mukaan (2004: 86) moniportainen relevanssiarviointi voi paljastaa vertailluista menetelmistä puolia, jotka eivät tule näkyviin binäärisessä relevanssiarvioinnissa. On esimerkiksi mahdollista, että kaksi menetelmää on suorituskyvyltään samankaltaisia binäärisellä asteikolla verrattaessa. Toinen menetelmä voi kuitenkin suoriutua paremmin erittäin relevanttien dokumenttien hakemisessa. Tässä työssä käytössä on aiherelevanssi. Relevanssiarviot ovat CLEF-aineistojen osalta binäärisiä ja Per Ahlgrenin kokoelman osalta moniportaisia (ks. luku 6.4.3).

2.5.2 Saanti ja tarkkuus

Kaksi tiedonhaussa yleisesti käytettyä haun onnistuneisuuden mittaria ovat saanti ja tarkkuus. Saanti ja tarkkuus perustuvat siihen, että hakutulos voidaan jakaa kahteen ryhmään, haussa löydettyihin ja haussa hylättyihin. Samoin relevanssiarvio voidaan jakaa kahtia relevantteihin ja epärelevantteihin. *Saannilla* tarkoitetaankin löydettyjen relevanttien dokumenttien suhdetta kaikkiin relevantteihin dokumentteihin. Saanti esitetään usein lukuna 0−1, ja luku kuvaa sitä, kuinka hyvin haussa pystytään löytämään tietokannan sisältämät relevantit dokumentit. Mitä suurempi luku on, sitä parempi saanti. *Tarkkuus* puolestaan tarkoittaa löydettyjen relevanttien dokumenttien suhdetta kaikkiin löydettyihin dokumentteihin. Myös tarkkuus esitetään lukuna 0−1, ja luku kuvaa sitä, kuinka suuri osa hakutuloksen dokumenteista on relevantteja. (Järvelin 1995: 55−56.)

Käytännön tiedonhauissa on kuitenkin mahdotonta arvioida saantia, koska se vaatisi relevanssiarviot miljoonista dokumenteista. Tällöin tiedossa olisi haun *absoluuttinen saanti*. *Suhteellisessa saannissa* käytetäänkin apuna jotakin tunnettua tietokannan sisältämien relevanttien dokumenttien joukkoa, johon haun saantia verrataan. (Järvelin 1995: 57.) Suhteellinen saanti voidaankin laskea saantikannan avulla. Saantikanta on otos dokumentteja, jotka ovat relevantteja. Saantikanta on muodostettu käsillä olevasta hausta riippumattomalla tavalla hakuaihetta analysoimalla ja hakuja suorittamalla. (Järvelin 1995: 64)

Saannin ja tarkkuuden esittäminen käytännössä tapahtuu *saanti-tarkkuus -käyrän* avulla. Tämä tarkoittaa sitä, että yksittäisten hakujen saanti- ja tarkkuusluvuista lasketaan keskimääräiset saanti- ja tarkkuusluvut, jotka esitetään käyränä. (Järvelin 1995: 60) Saanti-tarkkuus -käyriä käyte-

tään tavallisesti erilaisten hakualgoritmien suorituskyvyn vertailussa. Saanti-tarkkuus -käyrät mahdollistavat sekä tulosjoukon laadukkuuden että hakualgoritmin tehokkuuden määrällisen evaluoinnin. Toinen esitystapa on *DCV* eli *document cutoff value*, joka tarkoittaa keskimääräisen tarkkuuden laskemista tietyn suuruiselle tulosjoukolle, esimerkiksi kun 5, 10, 15, 20, 30, 50 tai 100 relevanttia dokumenttia on nähty. (Baeza-Yates & Ribeiro-Neto 1999: 78−79)

Tässä tutkimuksessa käytetään kahdesta edellä mainitusta esitystavasta saanti-tarkkuus -käyriä. Näiden lisäksi tiedonhaun tuloksia arvioidaan keskimääräisten tarkkuusarvojen avulla. *Tarkkuuksien keskiarvo* (MAP, mean average precision) on keskimääräinen ei-interpoloitu tarkkuuarvo kaikkien kyselyjen osalta. Tiedonhaun tutkimuksessa on kritisoitu saannin ja tarkkuuden käyttämistä mittareina ja esitetty vaihtoehtoisia mittaustapoja, mutta tässä tutkimuksessa käytetään mittareina perinteisesti saantia ja tarkkuutta.

2.6 Tiedonhaun laboratoriotutkimus

Tiedonhaun tutkimuksen lähestymistapoja ovat perinteinen eli järjestelmäkeskeinen tutkimus, käyttäjäkeskeinen tutkimus sekä uusin haara kognitiivinen tutkimus. Järjestelmäkeskeinen tutkimus keskittyy muun muassa tiedonhakutekniikoihin, tiedon esittämiseen, kontrolloituihin tieteellisiin testeihin sekä relevanssin tutkimiseen. Perinteiselle tutkimukselle läheisiä tieteenaloja ovat matematiikka, kielitiede ja tietojenkäsittely. Käyttäjäkeskeinen tutkimus on lisännyt vaikutusvaltaansa vuosien myötä ja siinä ollaan kiinnostuneita tiedonhakujärjestelmän käyttäjän käyttäytymisestä ja tiedontarpeista. Tutkimus keskittyykin tosielämän tutkimuksiin ja käyttäjämallinnukseen. Lähialoja ovat kognitiivinen psykologia, psykolingvistiikka ja sosiologia. Kolmas tutkimushaara on kognitiivinen tutkimus, joka keskittyy tiedonhakuprosessin kognitiiviseen puoleen eli tiedonhakijan kognitiiviseen tilaan. Lähialoja ovat kognitiiviset tieteet ja sosiologia. (Ingwersen 1992: 58)

Perinteistä tutkimushaaraa edustaa tiedonhaun laboratoriotutkimus, jolle on ominaista se, ettei käyttäjä osallistu tutkimukseen, vaan tiedonhakuprosessi on pitkälle automatisoitu ja hakuaiheet edustavat käyttäjiä. Laboratoriotutkimuksessa keskitytäänkin dokumenttien esityksiin, kyselyihin ja niiden täsmäyttämiseen toisiinsa. (Ingwersen & Järvelin 2005: 114–115) Perinteinen tiedonhaun tutkimus tutkii muun muassa sitä, miten dokumentit esitetään ja millaisia tiedonhakutekniikoita käytetään sekä millaisia mekaanisia osia tiedonlähteisiin ja tiedonhakujärjestelmiin sisäl-

tyy. Tavoitteena on hakutehokkuuden maksimoiminen, mihin pyritään vertailemalla erilaisia tekniikoita ja teorioita kontrolloidusti tietokannan tekstikokoelmien avulla. (Ingwersen 1992: 62)

Laboratoriomallin olennaisia osia ovat tietokanta, menetelmät, hakupyynnöt sekä relevanssiarviot. Relevanssia arvioidaan aiheenmukaisesti ja staattisesti. Kyse on siitä, miten dokumentin arvioija arvioi dokumentin vastaavan hakupyynnön aihetta. Oikeat käyttäjät eivät ole mukana arvioimassa relevanssia. (Ingwersen & Järvelin 2005: 4−5) Baeza-Yatesin ja Ribeiro-Neton (1999: 74) mukaan laboratoriotutkimusten vahvuuksina voidaan pitää toistettavuutta ja laajennettavuutta, vaikka tosielämän kokeet ovatkin kasvattaneet suosiotaan.

Robertsonin (1981: 11) mukaan tiedonhaun laboratoriotesti sisältää välttämättömiä osia. Ensiksikin tarvitaan tiedonhakujärjestelmä sääntöineen ja proseduureineen. Toiseksi tarvitaan aineistoa, jota järjestelmä käsittelee, eli dokumentteja ja hakupyyntöjä. Testien ja kokeiden tarkoituksena on vastata johonkin erityiseen kysymykseen, joten jokaisen testin olennainen osa koostuu koeasetelmasta. Tämän lisäksi kokeen tai testin toteuttaminen vaatii mittarin tai mittareita, joiden avulla saatuja tuloksia voidaan verrata ja analysoida. Mittareiden ohella tarvitaan menetelmiä tulosten analysointiin, jotta voidaan vetää johtopäätöksiä ja vastata koeasetelmassa asetettuihin tutkimuskysymyksiin.

Ingwersen ja Järvelin ovat kritisoineet perinteistä laboratoriomallia sen yksipuolisuuden, mekaanisuuden ja käyttäjän unohtamisen vuoksi. Laboratoriomallissa kiinnostuksen kohteena ovat dokumentin ja hakupyynnön esittäminen sekä näiden esitysten täsmäyttäminen toisiinsa. Todellisia käyttäjiä ja hakuaiheita ei sisällytetä malliin. Myöskin relevanssin arviointi on staattista ja aiheenmukaista. (Ingwersen & Järvelin 2005: 4–9.) Käyttäjä on kuitenkin aina jollain asteella mukana, koska tutkimusta tehdään käyttäjän toiminnan helpottamiseksi. Esimerkiksi luonnolliseen kieleen liittyvässä tutkimuksessa laboratoriotutkimus mahdollistaa kyseistä kielenilmiötä koskevan laajamittaisen testaamisen, jollaista käyttäjäsuuntautuneessa tutkimuksessa olisi vaikea toteuttaa.

3 LUONNOLLINEN KIELI TIEDONHAUSSA

Kuten aiemmin on jo todettu luonnollinen kieli on olennainen osa tekstitiedonhakua. Tiedonhaussa on osallisena kolme eri tasoa, joista yksi on luonnollisen kielen taso. Hakujärjestelmät eivät kuitenkaan osaa tulkita luonnollista kieltä, vaan ne toimivat merkkijonojen tasolla. Tässä luvussa esitellään luonnollisen kielen piirteitä nimenomaan siitä näkökulmasta, millaisia ongelmia ne aiheuttavat tiedonhaussa. Tutkimuksen kohteena on ruotsin kieli, joten esimerkit esittelevät pääasiassa ruotsin kieleen liittyviä piirteitä.

3.1 Kielen osajärjestelmät

Kieltä tarkastellaan osajärjestelmien kokonaisuutena. Tämän tutkielman näkökulmasta keskeinen kielen osajärjestelmä on *morfologia*, joka on sanojen sisäisen rakenteen osajärjestelmä. Toinen tutkielman aihepiiriä sivuava kielen osajärjestelmä on *semantiikka* eli merkitysten tutkimus. Muita osajärjestelmiä ovat fonologia eli äänneoppi, leksikko eli sanavarasto ja syntaksi eli lauseoppi. (Karlsson 2006: 15) Moniin kielen eri osajärjestelmiin liittyykin ilmiöitä, jotka aiheuttavat ongelmia tiedonhaussa. Se, millaiset ilmiöt tuottavat ongelmia, on myös kielikohtaista. Tämä tutkimus keskittyy luonnollisen kielen morfologisiin ilmiöihin. Tässä luvussa esittelen morfologiaan liittyviä kielen ilmiöitä, jotka aiheuttavat ongelmia tiedonhaun näkökulmasta. Lisäksi esittelen näiden ongelmien ratkaisemiseen kehitettyjä menetelmiä.

3.2 Morfologia

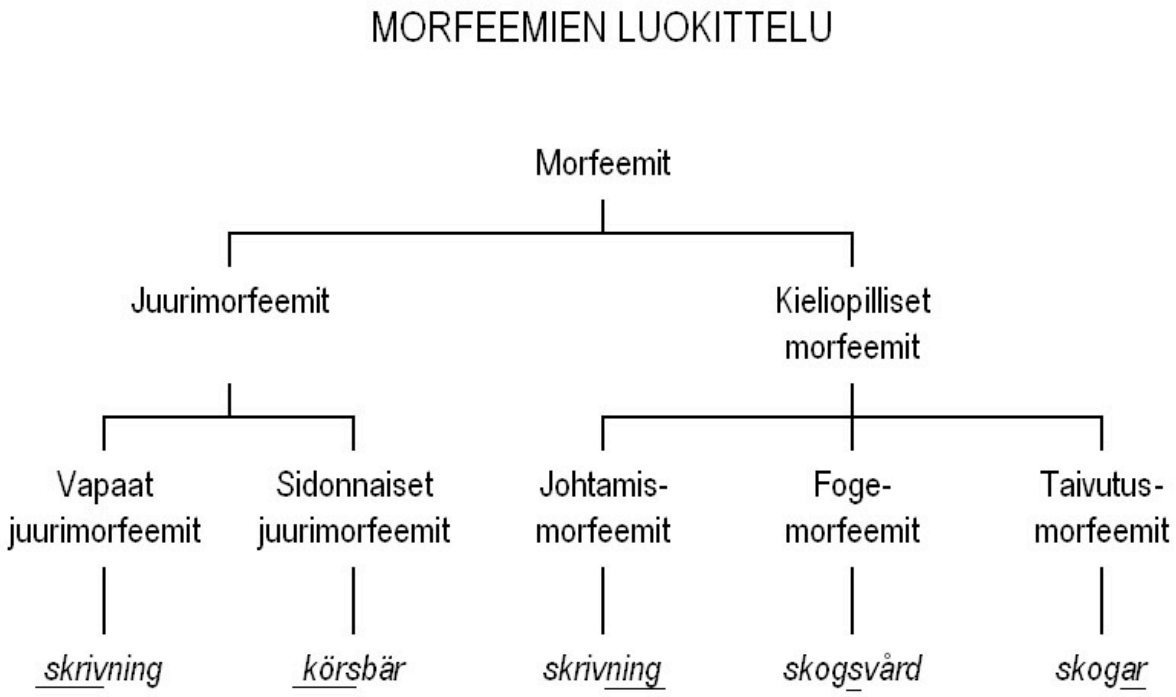
Seuraavaksi esittelen tarkemmin morfologian alaan kuuluvia kielen ilmiöitä eli sanojen taipumista ja sananmuodostusta.

3.2.1 Morfeemien luokittelu

Morfeemilla tarkoitetaan kielen pienintä merkitystä kantavaa yksikköä. Morfeemeja on vapaita ja sidonnaisia. *Vapaat morfeemit* voivat esiintyä itsenäisinä sanoina. *Sidonnaiset morfeemit* eivät voi esiintyä yksinään vaan ovat aina yhteydessä toiseen morfeemiin. (Liljestrand 1993: 20−22.) Häkkinen (2001: 115) jakaa vapaat morfeemit vielä kahtia. *Potentiaalisesti vapaat morfeemit* voivat esiintyä joko yksinään tai yhdessä muiden morfeemien kanssa (esimerkiksi **hus**, **husets**).

Vapaat morfeemit ovat Häkkisen mukaan puolestaan sellaisia, jotka eivät koskaan voi esiintyä yhdessä toisen morfeemin kanssa (esimerkiksi **och**, **men**). (Häkkinen 2001: 115−116)

Sidonnaiset morfeemit eli affiksit voidaan jakaa alaryhmiin sen mukaan, miten ne yhdistyvät kantasanaansa. Kantasanan edellä olevia morfeemeja ovat *prefiksit* eli etuliitteet (**på**-verka). Sanavartalon sisällä esiintyviä morfeemeja kutsutaan *infikseiksi* eli sisäliitteiksi. Niitä esiintyy erityisesti Euroopan ulkopuolisissa kielissä. Kantasanan jäljessä esiintyvät affiksit ovat puolestaan *suffikseja* eli jälki- tai loppuliitteitä (esimerkiksi mäkt-**ig**). Häkkisen mukaan potentiaalisesti vapaan morfeemin jakamatonta perusosaa voidaan puolestaan nimittää kannaksi, kantavartaloksi, (perus)vartaloksi tai juurimorfeemiksi. (Häkkinen 2001: 115−116)



KUVA 2. Morfeemien luokittelu Malmgrenia (1994: 28) mukaillen.

Kuvasta 2 käy ilmi Malmgrenin morfeemien luokittelu. Malmgren (1994: 26−28) jakaa morfeemit *juurimorfeemeihin* (rotmorfem) ja *kieliopillisiin morfeemeihin* (grammatiska morfem). Kieliopilliset morfeemit jakautuvat *taivutusmorfeemeihin* (böjningsmorfem, esimerkiksi skog-**ar**, kloc-**or**, hus-**en**), *johtamismorfeemeihin* (avledningsmorfem, skriv-**ning**, vän-**lig**, **för**-ändra) ja *fogemorfeemeihin* (skog-**s**-vård, kyrk-**o**-gård). Fogemorfeemit ovat sidosmorfeemeja, joilla yhdyssanan osat on liitetty yhteen (ks. luku 3.2.5). Juurimorfeemit voidaan jakaa vielä erikseen *vapaisiin juurimorfeemeihin* (**skriv**-ning) ja *sidonnaisiin juurimorfeemeihin* (**körs**-bär). (Malmgren 1994: 26−28)

Juurimorfeemi voi esiintyä itsenäisenä sanana tai pakollisessa yhdistelmässä toisen morfeemin kanssa. Esimerkki sidonnaisesta juurimorfeemista on *jäännösmorfeemi*. Jäännösmorfeemi on juurimorfeemi, joka voi esiintyä vain tietyssä yhdistelmässä ja jonka merkitys on kielenpuhujalle usein tuntematon (esimerkiksi **körs-**bär, **ling-**on). Toinen esimerkki sidonnaisesta juurimorfeemista on juurimorfeemi, joka on vierasta alkuperää (esimerkiksi **bom-**ull, **fru-**kost). (Liljestrand 1993: 22−23) Hultmanin mukaan sidonnaisilla juurimorfeemeilla on selkeä merkitys, mutta ne eivät voi esiintyä itsenäisinä sanoina, esimerkiksi morfeemi **multi** sanassa **multietnisk** sekä **mon(o)** sanoissa **monarki**, **monolog**. (Hultman 2003: 35)

3.2.2 Sanaan liittyvät käsitteet

Häkkisen (2001: 135) mukaan *sana* on kielen pienin vapaa muoto ja sen tunnusmerkkinä on, että se voi esiintyä itsenäisesti. Tiedonhaun näkökulmasta sanaksi voidaan ajatella kaikkia niitä kirjainjonoja, joita erottaa toisistaan välimerkki. Ruotsin kielestä löytyy tosin sanoja, jotka eivät täytä tätä määritelmää täydellisesti, esimerkiksi partikkeliverbit (**tycka om**, **omtyckt**). *Lekseemi* tarkoittaa sanaa sanaston tai sanakirjan yksikkönä. Lekseemin edustaja on siis usein *perusmuoto* tai *hakumuoto*. Häkkinen kuitenkin korostaa, että monissa sanaluokissa perus- ja hakumuodon määritteleminen voi tuottaa ongelmia. (Häkkinen 2001: 138.)

Sananmuoto on termi, jota voidaan käyttää lekseemin jostakin muodosta eli näin ollen myös tekstissä tai puheessa esiintyvistä sanoista. Tekstin rakennetta ja sanojen esiintymisfrekvenssiä tutkittaessa käytetään termiä *sanaesiintymä* eli *sane*. Tällöin tarkoitetaan erikseen jokaista tekstissä esiintyvää sananmuotoa. Koska sama lekseemi voi esiintyä eri muodoissa samassa tekstissä, on syytä pitää mielessä, lasketaanko lekseemejä, sananmuotoja vai esiintymiä. (Häkkinen 2001: 139)

Sanat voidaan Häkkisen mukaan jakaa rakenteensa perusteella kolmeen pääryhmään. Ydinsanaston muodostavat *jakamattomat perussanat* tai perusvartalot (suomessa **syödä**, **käsi**, **tehdä**). Toinen pääryhmä muodostuu johdetuista sanoista eli *johdoksista* (**käsi** – **käsittää**). Kolmas sananmuodostuksen rakennetyyppi on *yhdyssanat*, jotka sisältävät aina vähintään kaksi sanavartaloa (**kerrostalo**). (Häkkinen 2001: 141−142) Häkkisen mukaan sanojen määrän ilmoittaminen luonnollisessa kielessä on ongelmallista, koska sanasto muuttuu koko ajan. Sanoja jää pois käytöstä ja uusia muodostetaan lainaamalla, johtamalla ja yhdistämällä. *Leksikaalistunut sana* on muuttu-

nut yleisesti tunnetuksi ja käytetyksi osaksi sanastoa. Tällainen sana on kaikille kielen puhujille yhteinen, eikä sitä muodosteta joka kerta uudestaan. Selvimmin leksikaalistuneet sanat ovat semanttisesti ja morfologisesti läpinäkymättömiä. Niiden merkitys tai rakenne ei käy ilmi niistä itsestään. Toisaalta leksikaalistunut sana voi olla hyvinkin läpinäkyvä sana, joka on vakiintunut osaksi sanastoa. (Häkkinen 2001: 143)

3.2.3 Sanojen taipuminen

Kuten aikaisemmin esitellyistä morfeemityypeistä voi päätellä morfologian ilmiöt voidaan jakaa kahtia sanojen taipumiseen ja sanojen muodostamiseen. Sanojen muodostaminen puolestaan pitää sisällään sanojen johtamisen ja sanojen yhdistämisen. Suomen kielessä sanoja taivutetaan lisäämällä erilaisia päätteitä sanan vartalon perään. Yksi ryhmä affikseja ovatkin tunnukset, joiden avulla perusvartaloista muodostetaan alivartaloita (suomen kielessä monikon tunnus, esimerkiksi talo-**i**-ssa). Taivutuspäätteet seuraavat tunnuksia ja niitä ovat esimerkiksi nominien sijapäätteet (talo-**ssa**) ja verbien persoonapäätteet (soitta-**vat**). Viimeiseksi tulevat liitteet, suomen kielessä ensin omistusliitteet eli possessiivisuffiksit (talossa-**si**) ja sitten liitepartikkelit (talossasi-**kin**). (Häkkinen 2001: 116)

Kuvasta 3 käyvät ilmi ruotsin kielen taivutuskategoriat. Ruotsinkieliset substantiivit taipuvat neljässä morfologisessa kategoriassa: *luku*, *sija*, *spesies* ja *suku* (Karlsson 2006: 108). Ruotsin kielessä sanoilla on kaksi eri sukua, utrum (esimerkiksi **en bil**) ja neutrum (esimerkiksi **ett svin**). Lukua ilmaistaan yksiköllä ja monikolla ja monikon taivutus riippuu sanan suvusta (många **bilar**, många **svin**). Spesies viittaa sanan tunnettuuteen, jota ruotsin kielessä ilmaisevat epämääräinen (**en bil**, **bilar**) ja määräinen muoto (**bilen**, **bilarna**) sekä yksikössä että monikossa. Kieliopillisista sijoista ruotsin kielessä on käytössä nominatiivi eli perusmuoto (**svin**) ja genetiivi eli omistusmuoto (**svin-s**).

Spesies	Epämääräinen				
Suku		Utrum		Neutrum	
Luku		Yks	Mon	Yks	Mon
Sija	Nom	bil	bil+ar	svin	svin
	Gen	bil+s	bil+ar+s	svin+s	svin+s

Spesies	Määräinen				
Suku		Utrum		Neutrum	
Luku		Yks	Mon	Yks	Mon
Sija	Nom	bil+en	bil+ar+na	svin+et	svin+en
	Gen	bil+en+s	bil+ar+na+s	svin+et+s	svin+en+s

KUVA 3. Sanojen bil ’auto’ ja svin ’sika’ taivutusjärjestelmä (Karlsson 2006: 108–109).

Sanojen taipuminen aiheuttaa tiedonhaussa ongelmia siksi, että taivutusmuodot lisäävät kunkin sanan merkkijonovakioiden määrää, jolloin tiedonhaussa on huomioitava sanan kaikki mahdolliset kirjoitusasut ja taivutusvariantit. Ahlgren (2004: 42) kiinnittää myös huomiota ruotsin kielelle ominaiseen *mutaatioon*, joka tarkoittaa sanan vartalossa tapahtuvia vaihteluita. Esimerkiksi sanan **stad** monikkomuoto on **städer**. Jos sana **stad** katkaistaan kyselyssä, monikkomuotoa ei löydy. Toinen tiedonhaun kannalta olennainen piirre on vahvojen verbien esiintyminen ruotsin kielessä. Vahvoissa verbeissä niiden vartalossa tapahtuu taivutettaessa muutoksia. Karlsson (2006: 92) nimittää tätä ilmiötä *fuusioksi*. Esimerkiksi verbin **ligga** imperfektimuoto on **låg**, jota ei löydy katkaisemalla perusmuoto **ligga**.

3.2.4 Sanojen johtaminen

Taipumisen lisäksi morfologian alaan kuuluu myös sananmuodostus. Uusia sanoja syntyy kieleen muun muassa kahdella eri tavalla. Sanoja johdetaan ja sanoja yhdistetään toisiinsa. Sanojen johtamisessa *johtimiksi* kutsutaan affikseja, joiden avulla muodostetaan uusia sanoja. Johdettuja sanoja kutsutaan puolestaan *johdoksiksi*. Kun lähestytään asiaa sanojen johtamisen näkökulmasta, sanan perusosaa kutsutaan *kannaksi* tai *kantavartaloksi*, joka jää jäljelle, kun kaikki mahdolliset affiksit otetaan pois. Häkkisen mukaan johtaminen on rekursiivista, mikä tarkoittaa sitä, että samassa sanassa voi esiintyä monia johtimia peräkkäin. Eri johtimista muodostuu yksi kerrallaan

lisäämällä johdosketjuja. Termillä *kantasana* viitataankin siihen ketjun jäseneen, josta johdos on välittömästi muodostettu. (Häkkinen 2001: 141)

Hultmanin (2003: 33) mukaan ruotsin kielessä on noin kaksisataa eri johtamismorfeemia, joista osa on produktiivisia ja osa improduktiivisia. Produktiivisuus viittaa siihen, voiko johtimen avulla muodostaa uusia sanoja. Esimerkkejä produktiivisista johtimista ovat **are** (löpare), **eri** (slöseri), **het** (skönhet), **is** (kändis) ja **skap** (medlemskap), kun taas improduktiivisia johtimia nykypäivänä ovat **an** (väntan), **dom** (sjukdom) ja **else** (förståelse). (Hultman 2003: 55)

Varsinaisen johtamisen lisäksi yksi tapa muodostaa uusia sanoja on myös *takaperojohtaminen* (retrogradering, tillbakabildning). Näin esimerkiksi adjektiivin **vindsurfing** 'lainelautailu' avulla on muodostettu verbi **vindsurfa** 'lainelautailla', vaikka usein kehitys on kulkenut kielessä päinvastaiseen suuntaan (esimerkiksi verbeistä johdetut substantiivit). Toinen esimerkki on substantiivin **kedjerökare** 'ketjupolttaja' avulla muodostettu **kedjeröka** 'ketjupolttaa'. (Hultman 2003: 34) Johtamisen ja sanojen yhdistämisen lisäksi yksi merkittävä sananmuodostustapa on myös *lyhentäminen* lyhyiden sanojen (kortord) ja alkukirjainsanojen (initialord) muodossa. Esimerkiksi **automobil** on lyhentynyt muotoon **bil** ja **Dagens Nyheter** muotoon **DN**. (Hultman 2003: 34)

Sanojen johtaminen aiheuttaa tiedonhaussa ongelmia siksi, että sanan vartaloon liitettävät prefiksit ja suffiksit muuntavat kantasanan merkitystä ja saattavat piilottaa kantasanan kokonaan haussa (esimerkiksi **löpa** – **löpare**). On myös yleistä, että kantasanasta tehty johdos on leksikaalistunut niin paljon, että sen yhteys kantasanaan on lähes kokonaan hävinnyt (Ingwersen & Järvelin 2005: 151).

3.2.5 Sanojen yhdistäminen: yhdyssanat

Tutkielmassa kiinnostuksen kohteena ovat yhdyssanat. *Yhdyssanalla* tarkoitetaan ruotsin kielestä puhuttaessa yhtä yhteen kirjoitettua sanaa (vrt. esimerkiksi englanti: **information retrieval**). Yhdyssana jakautuu vähintään kahteen sanamaiseen pääosaan, joista kumpikin voi myös toimia itsenäisenä sanana (Malmgren 1994: 32). On olemassa monia perusteita sille, miksi yhdyssanan loppuosa on alkuosaa tärkeämpi osa. Yhdyssanan loppuosa määrittää yhdyssanan sanaluokan. Yhdyssana on usein myös loppuosansa *hyperonyymi*. Esimerkiksi sanassa **rödvin** punaviini on yhdenlainen viini ja siihen voidaan viitata tekstissä pelkästään sen loppuosalla viini. (Malmgren 1994: 34–35.) Yhdyssanojen esiintyminen kielessä on tärkeä ilmiö tiedonhaun näkökulmasta,

koska ne ja erityisesti niiden loppuosat ovat usein merkitystä kantavia sanoja niin kyselyissä kuin dokumenteissakin (Hedlund 2002: luku 1).

Hedlund (2002) on huomioinut yhdyssanojen merkityksen tiedonhaussa ja käsittelee väitöskirjaansa liittyvässä artikkelissa aihetta kieltenvälisen tiedonhaun näkökulmasta. Hedlundin (2002: luku 3) tekemän testin mukaan yhdyssanojen yleisyys 100 000 sanan sanomalehtiaineistossa on suomen kielessä 8,7 prosenttia, ruotsin kielessä 9,8 prosenttia ja saksan kielessä 10,2 prosenttia. Aineistona on käytetty CLEF2000- ja 2001-kokoelmien erikielisistä aineistoista poimittua otosta. Tutkielman tarkoituksena onkin täydentää ruotsin kieltä koskevaa yhdyssanojen esiintymien kartoittamista ja tutkia yhdyssanojen esiintymistä hakuaiheissa ja dokumenteissa. Yhdyssanojen taajuuksia on aikaisemmin esitetty muun muassa Allénin ja muiden (1980) frekvenssisanakirjassa. Sanakirja ei tosin tarjoa kaikkia yhdyssanoja kattavaa tietoa, vaan sanakirjassa on lueteltu eri yhdyssanojen esiintymisfrekvenssit.

Hedlundin mukaan yhdyssanojen käsittelymenetelmillä voi olla suurta vaikutusta tiedonhakuun, koska varsinkin kieltenvälisessä tiedonhaussa käännössanakirjat eivät usein sisällä kuin kaikista yleisimmät yhdyssanat. Näin ollen esimerkiksi jakamalla yhdyssanat osiin voidaan tehdä käännöksiä myös sellaisista sanoista, joita ei muuten löytyisi sanakirjoista (Hedlund 2003: 18). Uusia yhdyssanoja syntyy kieleen produktiivisesti. Monet yhdyssanat ovat niin sanottuja tilapäisiä yhdyssanoja, jotka muodostetaan vain tilapäistä käyttöä varten ja joita ei siksi löydy sanakirjasta (Hedlund 2002: luku 2). Näin ollen niiden osittaminen voi olla hyödyllistä. Yksikielisessäkin tiedonhaussa tällä on merkitystä. Jos yhdyssanoja ei jaeta osiin, yhdyssanojen loppuosat jäävät haussa huomioimatta. Yleensä nämä loppuosat ovat kuitenkin yhdyssanojen tärkeimpiä sanoja ja siksi tärkeitä hakuavaimia (Hedlund 2002: luku 1).

Yksi yhdyssanoihin liittyvä ruotsin kielen erityispiirre on fogemorfeemien esiintyminen yhdyssanojen välissä. Ne ovat liitännäisiä, joilla yhdyssanan osat on liitetty yhteen. Hedlund ja muut (2001: 153) esittävät, että morfologisen analyysiohjelman tulisi pystyä huomioimaan fogemorfeemien esiintyminen kielessä. Esimerkiksi yhdyssana **skogsindustrin** on perusmuotoistamisen ja yhdyssanojen osittamisen jälkeen muotoa **skogs-industri**, jolloin yhdyssanan alkuosa on **skogs** eikä pelkällä muodolla **skog** löydetä mitään. Ruotsin kielen morfologinen analyysiohjelma SWETWOL pystyy pääosin tunnistamaan yhdyssanojen osien välissä olevat liitännäiset, koska ohjelma palauttaa myös yhdyssanojen osat perusmuotoonsa. Lisää SWETWOL:n toiminnasta on luvassa luvussa 3.5.5.

3.2.6 Yhdyssanojen luokittelu

Yhdyssanat voivat koostua monen eri sanaluokan sanoista. Erilaisia yhdistelmävaihtoehtoja on paljon. Yksi tapa luokitella yhdyssanoja on jako nominaalisiin yhdyssanoihin, johdettuihin yhdyssanoihin ja yhdyssanaverbeihin. *Nominaaliset yhdyssanat* tarkoittavat substantiiviin (**valbudget**) tai adjektiiviin (**miljövänlig**) päättyviä yhdyssanoja. (Liljestrand 1993: 39.) *Johdetut yhdyssanat* (avledda sammansättningar) sijoittuvat yhdyssanojen ja johdoksien välimaastoon. Johdetun yhdyssanan etuosa on itsenäinen sana ja loppuosa johdos. Yhdyssanan loppuosaa ei käytetä itsenäisesti, vaan se vaatii etuosan ollakseen ymmärrettävä. Esimerkkejä johdetuista yhdyssanoista ovat sanat **svart-ögd** ja **mot-strävig**. **Ögd** ja **strävig** eivät sellaisinaan voi toimia itsenäisinä sanoina. (Liljestrand 1993: 47.) Esimerkiksi SWETWOL kuitenkin pitää mainittuja esimerkkisanoja pikemminkin yhdyssanoina kuin johdoksina. Kolmas tyyppi käsittää *yhdyssanaverbit* (sammansatta verb). Niissä yhdyssanan etuosa voi olla jotakin muuta sanaluokkaa kuin verbi, esimerkiksi substantiivi (**lag-stifta**), prepositio (**fram-ställa**) tai adjektiivi (**enkel-rikta**). Myös verbi etuosana on mahdollinen (**stört-dyka**). (Liljestrand 1993: 50.) Enemmistö yhdyssanoista on muodostettu substantiiveista. Substantiivisten yhdyssanojen voisi ajatella kantavan paljon merkityksiä ja ne ovat siksi tärkeitä tiedonhaun näkökulmasta (Hedlund 2002: luku 2). Tutkielman tarkoituksena on myös selvittää yhdyssanojen sanaluokkien jakautumista aineistossa.

Yhdyssanoja voidaan myös luokitella niiden merkityksen mukaan. Yksi tapa jaotella on jako determinatiivisiin ja kopulatiivisiin yhdyssanoihin. *Determinatiivisen yhdyssanan* loppuosa on pääsana, jota etuosa määrittää (esimerkiksi **expresståg**). *Kopulatiivisen yhdyssanan* etu- ja jälkiosa ovat keskenään samanarvoisia (esimerkiksi **blågul**). (Liljestrand 1993: 44−47.) Toinen tärkeä erottelu on kompositionaalisten ja ei-kompositionaalisten yhdyssanojen välillä. *Kompositionaalisen yhdyssanan* merkitys on läpinäkyvä eli transparentti. Merkitys siis käy ilmi yhdyssanan osista (esimerkiksi **sköljvatten**). Osa tällaisista sanoista on muodostettu vain tilapäistä käyttöä varten. *Ei-kompositionaalisen yhdyssanan* merkitys on läpinäkymätön eli opaakki. Tämä tarkoittaa sitä, ettei yhdyssanan kokonaismerkitys seuraa osien merkityksistä (esimerkiksi **jordgubbe**). Tällaiset yhdyssanat kuuluvat usein perussanastoon ja niiden merkitys on myös leksikaalistunut. (Karlsson 2006: 193; Malmgren 1994: 24.) Leksikaalistuminen tarkoittaa sitä, että sana esiintyy sanakirjassa ja se on vakiintunut osaksi kielen perussanastoa.

On tapauksia, joissa kompositionaalisenkaan yhdyssanan täsmällinen merkitys ei ole tulkittavissa sen osista. Pirkolan (1999: 19) antaman esimerkin mukaan **paperipussi** on pussi, joka on tehty paperista. **Paperikone** puolestaan on kone, joka valmistaa paperia. **Paperikoivu** ei ole valmistettu paperista. Kaikki nämä yhdyssanat ovat loppuosiensa *hyponyymejä* eli suppeampia termejä. Pirkolan mukaan hyponyymi-hyperonyymi –suhde onkin tyypillistä kompositionaalisille yhdyssanoille.

Pirkola (2001: 343) esittää myös, että olisi tarpeen tutkia kompositionaalisten yhdyssanojen määrää ruotsin kielessä. Näin voitaisiin tehdä päätelmiä siitä, miten ruotsin kieltä kannattaa käsitellä tiedonhaussa. Jos suurin osa yhdyssanoista on kompositionaalisia, voidaan olettaa, että yhdyssanojen osittaminen kannattaa ruotsinkielisessä tiedonhaussa. Ei-kompositionaalisten yhdyssanojen osittaminen ei ole kannattavaa, koska niiden muodostamaa kokonaisuutta ei ole tarpeen rikkoa osiin. Pirkola esittää myös kertoimen laskemista eri kielille kompositionaalisten yhdyssanojen määrän osalta. Kertoimen avulla kielten välinen vertailu olisi helpompaa ja tutkimustuloksiakin voitaisiin soveltaa paremmin kielestä toiseen. (Pirkola 2001: 343.)

Olli Blåbergin (1988) kielitieteellinen korpustutkimus kartoittaa ruotsinkielisten yhdyssanojen syntaktista ja semanttista luonnetta. Tutkimuksessa on myös selvitetty, kuinka suuri osa ruotsin kielen yhdyssanoista on leksikaalistuneita (vrt. ei-kompositionaaliset yhdyssanat). Tutkimuksessa todetaan, että leksikaalistumista esiintyy substantiivisia yhdyssanoja enemmän adjektiivisissa ja verbaaleissa yhdyssanoissa (Blåberg 1988: 53). Tutkimuksesta on kuitenkin vierähtänyt jo aikaa, joten aihetta olisi syytä tutkia lisää nykyruotsin ja nimenomaan tiedonhaun näkökulmasta. Tutkielmassa pyritäänkin kartoittamaan ruotsin kielessä esiintyviä yhdyssanatyyppejä. Jaotteluna käytetään yhtäältä sanaluokkajakoa ja toisaalta jakoa kompositionaalisiin ja ei-kompositionaalisiin yhdyssanoihin.

Hellbergin (1978: 22) mukaan ruotsin kieli tarjoaa rajoittamattomat mahdollisuudet muodostaa tilapäisiä yhdyssanoja. Tämä seikka pienentää ruotsin kieltä varten laadittavan sanakirjan kokoa, koska tilapäisiä yhdyssanoja ei kannata tallentaa sanakirjaan. Ruotsin kielen sanakirjoissa on Hellbergin mukaan kahdenlaisia yhdyssanoja ja johdoksia. Ensimmäinen tyyppi on rakentunut muodollisesti ja epäsäännöllisesti. Esimerkiksi yhdyssana **lantbruk** tulee sanasta **land**, **smörgås** (monikossa **smörgåsar**) sanasta **gås** (monikossa **gäss**). Toisessa tyypissä yhdyssanan osat on helppo tunnistaa (esimerkiksi **trädgård**). Jako muistuttaa siis edellä esiteltyä jaottelua kompositionaalisiin ja ei-kompositionaalisiin yhdyssanoihin. Hellbergin mukaan sanakirjaa muodostetta-

essa olisi hyvä muistaa, että sellaisten yhdyssanojen tai johdosten, joiden merkitystä ei voi päätellä niiden osista, pitäisi olla sellaisinaan hakuavaimina sanakirjassa. (Hellberg 1978: 22)

Ruotsin kielessä on myös yleistä moniosaisten yhdyssanojen muodostaminen, jota Hellberg (1978: 25−26) kutsuu nimellä ”double compounding”. Moniosaisuus myös vaikuttaa sanan vartaloon. Esimerkiksi sana **nässla** on kaksiosaisessa yhdyssanassa muotoa **nässelblad**, moniosaisessa puolestaan muodossa **brännässleblad**. Ruotsin kielessä on myös paljon yhdyssanoja, joiden etuosa ei voi toimia itsenäisenä sanana. Esimerkiksi yhdyssanoissa **justitiedepartement**, **generalförsamling**, **psykoanalys** ja **svärfar**. (Hellberg 1978: 26−27) Tiedonhaussa moniosaisuus vaikuttaa siten, että yhdyssanojen osittamisen myötä voi syntyä hakuavaimia, jotka eivät ole oikeassa muodossa. Tämä saattaa lisätä hakuavainten monitulkintaisuutta tiedonhaussa.

Yhdyssanojen osittamisen näkökulmasta yhdyssanoja voidaan tarkastella kolmesta eri näkökulmasta. Ensimmäinen yhdyssanatyyppi on pitkä tilapäisyhdyssana (esimerkiksi **flygplatsolycka**), joka ei ole vakiintunut. Tämän vuoksi yhdyssanan osittaminen on olennaista, koska yhdyssana ei välttämättä esiinny sellaisenaan kaikissa teksteissä. Toinen tyyppi on vakiintunut ja merkitykseltään läpinäkyvä yhdyssana (esimerkiksi **vindkraft**). Tästä tyypistä osa yhdyssanoista on luonteeltaan sellaisia, että osittaminen kannattaa, mutta joukossa on myös yhdyssanoja, joita ei välttämättä ole hyödyllistä jakaa osiin (esimerkiksi **regelverk**). Kolmas tyyppi on vakiintunut ja merkitykseltään läpinäkymätön yhdyssana (esimerkiksi **jordgubbe**), jonka osittaminen on harvoin tarkoituksenmukaista.

3.3 Semantiikka: synonymia, homografia ja polysemia

Semantiikan eli merkitysopin ilmiöt sivuavat tämän tutkielman aihepiiriä. Semantiikan osalta tiedonhaussa ongelmia tuottavat synonymia, homografia ja polysemia. *Synonymialla* viitataan siihen, että samasta käsitteestä voidaan käyttää monia eri ilmaisuja, mikä tiedonhakijan olisi huomioitava haussa käyttämällä rinnakkaisia hakuavaimia. *Homografia* on kirjoitusasuun liittyvää monitulkintaisuutta. Yli 65 prosenttia ruotsinkielisen tekstin sananmuodoista on homografeja. Esimerkiksi ruotsin kielen sana **en** voi olla artikkeli, lukusana, substantiivi (’kataja’), pronomini (**Om man ser en, måste man akta sig** ’Jos joku näkee sinut, täytyy olla varovainen’) ja adverbi (**en 3–4 gånger** ’noin 3–4 kertaa’). Myös ruotsin kielen sanalla **för** on Karlssonin mukaan kahdeksan eri kieliopillisen sanan ilmentymää. Suomen kielessä vastaava homografien esiintymistiheys on vain noin 15 prosenttia. (Karlsson 2006: 88–89.) *Polysemialla* puolestaan

tarkoitetaan sanojen monimerkityksisyyttä eli samalla sanalla voidaan viitata useampaan eri tarkoitteeseen (Karlsson 2006: 213). Esimerkiksi sana **stjärna** voi viitata taivaalla loistavaan tähteen tai konsertissa esiintyvään tähteen.

3.4 Ruotsin kielen keskeiset piirteet tiedonhaun näkökulmasta

Edellä on esitelty morfologian ja semantiikan ilmiöiden ongelmallisia piirteitä tiedonhaun näkökulmasta. Ruotsin kielen piirteitä on käsitelty tiedonhaun näkökulmasta Hedlundin, Pirkolan ja Järvelinin (2001) artikkelissa. Ruotsin kieli on tiedonhaun kannalta haastava kieli muun muassa morfologisten piirteidensä ansiosta. Hedlundin ja muiden (2001: 151–154) mukaan ruotsin kielessä haasteita aiheuttavatkin seuraavat piirteet:

1. Morfologian piirteet
2. Suvun ilmaisemiseen liittyvät seikat
3. Sanojen johtaminen
4. Yhdyssanojen runsas esiintyminen
5. Homografisten sanojen runsas esiintyminen

3.5 Kielenkäsittelymenetelmät

Edellä on esitelty pääasiassa morfologian mutta myös semantiikan ilmiöitä luonnollisessa kielessä ja tiedonhaussa. Kielet vaihtelevat paljon morfologisilta ominaisuuksiltaan. Kielissä, joiden sananmuodoissa on vähän morfologista variaatiota, morfologisia ongelmia ja sanojen morfologista käsittelyä ei tarvitse ottaa huomioon (Pirkola 2001: 341). On kuitenkin myös olemassa kieliä, joissa tällaista vaihtelua on paljon. Näissä kielissä morfologia aiheuttaa ongelmia sen vuoksi, että hakuavaimina ja hakemistosanoina esiintyvät sanat vaihtelevat muodoiltaan paljon. Morfologian aiheuttamat ongelmalliset piirteet voidaan siis jakaa kolmeen alueeseen, joihin kaikkiin liittyy sananmuotojen morfologinen variaatio: sanojen taipuminen, sanojen johtaminen ja yhdyssanat (Pirkola 2001: 332). Dokumentit eivät löydy, elleivät hakemistosana ja hakuavain vastaa toisiaan täydellisesti. Jos dokumenttien tallennusvaiheessa käytetään taivutusmuotoista hakemistoa, tiedonhakijan on osattava katkaista käyttämänsä hakuavaimet kyselyissä. Katkaisu toimii useimmissa kielissä, mutta esimerkiksi suomen kielessä on sanoja, joita on mahdotonta katkaista (esimerkiksi **yö** – **öiden** , **työ** – **töiden**). Luonnollisen kielen moninaisuus aiheuttaa myös tiedonhakijalle suuria ongelmia, koska hänen pitää kiinnittää erityistä huomiota sanojen muotoon.

Kaikki tiedonhakijat eivät ole yhtä kokeneita eivätkä välttämättä ole tottuneet katkaisemaan sanoja haussa.

Eräänlainen sateenvarjokäsite tutkielmassani on *luonnollisen kielen käsittely* (natural language processing, NLP), joka pyrkii tarjoamaan ratkaisuja luonnollisen kielen ongelmakohtiin. Luonnollisen kielen käsittelyllä tarkoitetaan tietoteknisten menetelmien kehittämistä ja soveltamista luonnolliseen kieleen (Carlson & Honkela 1993: 233). Menetelmiä voidaan kehittää kielen eri tasojen ongelmien ratkaisemiseen. Morfologisista ongelmista sananmuotojen morfologinen variaatio voidaan ottaa huomioon monella eri tavalla. Sanoja voidaan katkaista haettaessa tai voidaan käyttää jokerimerkkejä joidenkin merkkijonojen korvaajina. Nämä ratkaisut vaativat kuitenkin käyttäjältä paljon. Käyttäjän toimintaa helpottamaan onkin kehitetty kielenkäsittelymenetelmiä, joita voidaan hyödyntää dokumenttien tallennusvaiheessa muodostamalla erilaisia hakemistoja. Kielenkäsittelymenetelmiä on myös mahdollista kytkeä mukaan hakukoneen toimintaan. Tässä tutkimuksessa kielenkäsittelymenetelmiä lähestytään erityisesti erilaisten kyselynmuodostusmenetelmien näkökulmasta.

3.5.1 Perusmuotoistaminen

Perusmuotoistaminen tarkoittaa sananmuotojen palauttamista perusmuotoon (esimerkiksi **fotbollslagets** saa muodon **fotbollslag**). Menetelmä perustuu kontekstiriippumattomaan yksittäisten sananmuotojen analyysiin sanakirjan ja säännöstön avulla. Analyysin suorittaa morfologinen analyysiohjelma. Kun dokumenttien sisältämät sanat sijoitetaan perusmuotoisina dokumentin hakemistoon, käyttäjä voi käyttää hakiessaan perusmuotoisia sanoja eikä hänen tarvitse ottaa huomioon sanojen taipumista. Perusmuotoistamismenetelmälle on ominaista se, että se tuottaa kaikki mahdolliset tulkintavaihtoehdot. Koska sananmuotojen analyysissa ei oteta huomioon niiden esiintymiskontekstia, homografia aiheuttaa sananmuotojen monitulkintaisuutta. (Järvelin 1995: 171.)

Kontekstin kannalta virheellisten tulkintojen tuottamista kutsutaan *ylitulkinnaksi.* Tämä ylitulkinta vaikuttaa tiedonhaun tehokkuuteen, koska se vähentää hakujen tarkkuutta. Ylitulkinnan vuoksi onkin kehitetty toinen menetelmä *disambiguointi*, joka pyrkii ottamaan huomioon sanojen muotojen lisäksi myös niiden syntaktisen aseman lauseessa. (Järvelin 1995: 171.) Tutkimukset (Leppänen 1995; Sanderson 1996) ovat kuitenkin osoittaneet, että disambiguoinnilla ei ainakaan vielä toistaiseksi saavuteta suuria etuja.

3.5.2 Yhdyssanojen käsittely

Perusmuotoistamisen yhteydessä on myös mahdollista käsitellä yhdyssanoja. Molemmat käsittelyt on mahdollista toteuttaa morfologisella analyysiohjelmalla. *Yhdyssanojen osittaminen* tarkoittaa yhdyssanojen jakamista osiin (esimerkiksi **fotbollsmatch** on ositettuna **boll**, **fotboll**, **bolls**, **fot**, **match**, **fotbolls**, **fotbollsmatch**). Yhdyssanojen automaattinen osittaminen tapahtuu morfologisen analyysiohjelman avulla ilman ihmisen puuttumista sen toimintaan. Yhdyssanat voidaan kuitenkin osittaa myös valikoivasti, jolloin voidaan intellektuaalisesti valita, mitkä yhdyssanat halutaan osittaa analyysiohjelmalla.

Yhdyssanojen eliminointiperiaate puolestaan tarkoittaa sitä, että useampiosaisen yhdyssanan yhdysosista otetaan mukaan vain minimimäärä (esimerkiksi **fotbollsmatch** saa muodot **fotboll**, **match**, **fotbolls**, **fotbollsmatch**). Tämäkin yhdyssanojen käsittelymenetelmä on siis valikoivaa, morfologisen analyysiohjelman suorittamaa toimintaa. Päätavoitteena on valita pienin tulkintojen määrä. Esimerkiksi yhdyssana **tietokannastakin** voidaan tulkita monella eri tavalla: **tie-tokannastakin**, **tieto-kannas-takin**, **tieto-kannastakin**. Jotta tiedonhaku onnistuisi, oikea tulkinta on viimeinen, joka erottaa yhdyssanan kaksi eri osaa **tieto-kanta**. Yhdyssanojen eliminointiperiaate auttaa tunnistamaan olennaiset sanat. Yhdyssanaraja voisi olla monessa eri kohdassa, mutta mukaan valitaan vain se yhdyssanaraja, joka tuottaa pienimmän tulkintojen määrän.

3.5.3 Karsinta-algoritmit

Edellä esiteltyjen kielenkäsittelymenetelmien lisäksi on mahdollista hyödyntää stemmausta eli karsintaa (stemming). *Karsinta* on kielenkäsittelymenetelmä, joka karsii sananmuodoista niiden pääteainekset siten, että jäljelle jää stemmi eli vartalo, joka ei kuitenkaan välttämättä ole enää ymmärrettävä sana. Perusmuotoistamisen ja karsinnan ero onkin juuri käsittelyn lopputulos. Esimerkiksi sananmuodot **cykel** ja **cyklade** palautetaan karsinnassa samaan muotoon **cykl** (Carlberger et al. 2001: 2). Stemmi eli vartalo on se osa sanasta, joka jää jäljelle affiksien poistamisen jälkeen. Esimerkiksi **connect** on karsintavartalo sananmuodoista **connected**, **connecting**, **connections** ja **connection**. Karsintaa voidaan pitää hyödyllisenä, koska se vähentää samasta juurimorfeemista muodostettavien varianttien määrää ja kehittää niille yhteisen käsitteen. Karsinnan on myös todettu pienentävän indeksointirakenteen kokoa, koska eri hakemistosanojen määrä pienentyy karsimisen ansiosta. Eri tutkimukset antavat kuitenkin varsin ristiriitaisia tuloksia siitä,

onko karsinta todella hyödyllistä. Myös kielten välillä on suuria eroja tämän suhteen. (Baeza-Yates & Ribeiro-Neto 1999: 168)

3.5.4 Ruotsin kielen morfologinen käsittely

Ruotsin kielen erityispiirteet voidaan ottaa monin eri tavoin huomioon tiedonhaussa. Ratkaisut ovat pitkälti edellä esitellyn kaltaisia. Morfologian osa-alueisiin ratkaisu löytyy perusmuotoistamisesta ja yhdyssanojen käsittelystä, joskin Hedlund ja muut (2001: 147) toteavat, että morfologisen analyysiohjelman toimintaan liittyy paljon heikkouksia. Ruotsin kieltä varten on myös käytettävissä karsinta-algoritmi eli stemmeri. Ruotsin kielessä yleiset homografiatapaukset pystytään erottamaan, kun nähdään sananmuodon konteksti. Kun on kyse automaattisesta kielenkäsittelystä kuten perusmuotoistamisesta, konteksti ei ole näkyvissä, vaan sana saa kaikki mahdolliset tulkintavaihtoehdot. Sananmuotojen disambiguointi ottaa huomioon myös sananmuodon syntaktisen aseman, joten tällä menetelmällä on mahdollista ratkaista homografiaan liittyviä ongelmia. Monitulkintaisuuteen ratkaisun voi tarjota myös *sanaluokkien tunnistaminen* (POS, part-of-speech tagging) (Strzalkowski et al. 1999: 124). Sanaluokkien tunnistamisella esimerkiksi för-prepositio voitaisiin erottaa föra-verbistä (Hedlund et al. 2001: 158). Vaikkakin edellä mainituista menetelmistä on jo todettu olevan apua kielellisten ongelmien ratkomisessa, lisää tutkimusta ja kehittämistä kaivataan yhä (Strzalkowski et al. 1999: 143).

3.5.5 Morfologinen analyysiohjelma: Esimerkkinä SWETWOL

Perusmuotoistaminen ja yhdyssanojen käsittely voidaan toteuttaa *morfologisella analyysiohjelmalla*, jonka toiminta perustuu kontekstiriippumattoman sanakirjan ja kieliopillisten sääntöjen käyttöön. Ruotsin kielessä perusmuotoistaminen ja yhdyssanojen osittaminen on mahdollista toteuttaa SWETWOL-analyysiohjelmalla, jonka toimintaa on esitellyt muun muassa Fred Karlsson (1992). SWETWOL:n toimintaperiaate perustuu Kimmo Koskenniemen (1983) kehittämään kaksitasomalliin. Samaa mallia on sovellettu myös muiden kielten analyysiohjelmien kehittämisessä.

SWETWOL-ohjelman toiminta perustuu sanakirjan ja taivutussääntöjen käyttöön. SWETWOL:n kuten muidenkin analyysiohjelmien ongelmana on ylitulkinta. Koska ohjelma keskittyy sananmuotojen morfologiseen analyysiin ottamatta huomioon niiden esiintymiskontekstia, sananmuotojen monitulkintaisuus aiheuttaa ongelmia. Sananmuoto tulkitaan monitulkintaiseksi, jos

SWETWOL antaa siitä enemmän kuin yhden tulkinnan (Karlsson 1992: 29). Seuraavassa on ohjelman tekemät analyysit sanoista **dinosaurielämning** ja **arbetsförhållandena**. **Dinosaurielämning** on esimerkki yksitulkintaisesta sanasta:

```
"<dinosaurielämning>"
        "dinosaurie#lämning"  N UTR INDEF SG NOM
```

Yhdyssanojen osien väliin ohjelma sijoittaa merkin #. **Arbetsförhållandena** on taivutusmuodossa oleva sana, jonka ohjelma ensin perusmuotoistaa väärin ja sen vuoksi myöskin jakaa yhdyssanan vääristä kohdista osiin. Sana on siis selvästi monitulkintainen:

```
"<arbetsförhållandena>"
        "arbetsför#hållande"  N NEU DEF PL NOM
        "arbetsför#hål#land"  <CLLQ> N NEU DEF PL NOM
        "arbetsför#håll#land"  <CLLQ> N NEU DEF PL NOM
        "arbets#förhållande"  <RETAIN!> N NEU DEF PL NOM
        "arbets#förhållande"  N NEU DEF PL NOM
```

4 MORFOLOGINEN KÄSITTELY TUTKIMUSKOHTEENA

Kielten morfologista käsittelyä on tutkittu eri kielissä niin kielitieteen kuin tiedonhaunkin näkökulmasta. Tutkituin kieli tiedonhaun näkökulmasta on englanti. Englannin kielen perusteella tehdyt päätelmät eivät kuitenkaan ole yksinään riittäviä, koska kielet eroavat toisistaan monin tavoin. Esimerkiksi englannin kieli suosii sanaliittoja, ruotsissa puolestaan muodostetaan pikemminkin yhdyssanoja (Hedlund 2002: luku 2). Ruotsin kielessä yhdyssanalla tarkoitetaankin vain yhtä sanaa (esimerkiksi **informationsåtervinning**), kun taas englannin kielessä yhdyssanoihin voidaan laskea kuuluviksi sanaliitotkin (esimerkiksi **information retrieval**). Näin ollen englannin kielen perusteella tehdyt päätelmät ja englannin kieltä varten suunnitellut kielenkäsittelyohjelmat eivät riitä.

Tässä luvussa esittelen kielten morfologista käsittelyä tutkimuskohteena. Erityisesti keskityn ruotsin kieltä käsitteleviin tutkimuksiin, vaikkakin ruotsin kieltä on tiedonhaun näkökulmasta tutkittu suhteellisen vähän. Tutkimuksia aiheesta on tehty kahdesta näkökulmasta. Ensiksikin on tarkasteltu ruotsin kielen ominaispiirteitä tiedonhaun kannalta. Toiseksi on vertailtu erilaisten tallennus- ja hakumenetelmien tehokkuutta tiedonhaussa ruotsin kielessä. Tässä luvussa esittelen lähinnä jälkimmäisen tutkimushaaran keskeisimpiä tutkimustuloksia.

Per Ahlgren (2004) on tutkinut väitöskirjassaan erilaisten hakemistojen ja hakuavainten yhdistelmien vaikutuksia hakutehokkuuteen ruotsin kielessä. Vertailukohtana (baseline) on tekstissä esiintyvien sananmuotojen käyttäminen sellaisinaan. Vertailtavia hakemistoja ovat taivutusmuotoinen hakemisto, perusmuotoistettu hakemisto, jossa yhdyssanat ovat osittamatta, perusmuotoistettu hakemisto, jossa yhdyssanat on jaettu osiin sekä perusmuotoinen hakemisto, jonka yhdyssanat on jaettu osiin ja jossa on lisäksi hyödynnetty yhdyssanojen eliminointiperiaatetta (Ahlgren 2004: 65–66). Hakemistoista on haettu eri muodoissa olevilla hakuavaimilla siten, että taivutusmuotoisen hakemiston kanssa on käytetty katkaistuja hakuavaimia ja perusmuotoistettujen hakemistojen kanssa perusmuotoisia hakuavaimia. Parhaiten vertailussa menestyy taivutusmuotoisen hakemiston ja katkaistujen hakuavainten yhdistelmä (Ahlgren 2004: 102).

Yhdyssanojen osittamista kyselyissä ei kuitenkaan ole suoritettu. Yhdyssanat on siis ositettu tallennusvaiheessa, mutta hakuvaiheessa perusmuotoistettujen hakemistojen yhteydessä on käytetty perusmuotoistettuja osittamattomia hakuavaimia. Ahlgren toteaakin, että olisi hyödyllistä testata sellaista hakumenetelmää, jossa yhdyssanat on sekä perusmuotoistettu että jaettu osiin. Ahlgren kuitenkin arvioi, että osittamisen vaikutus hakujen tehokkuuteen riippuu kyselyn aiheesta ja hakuavaimista eli tehokkuus on kyselykohtaista. (Ahlgren 2004: 131–133.)

Eija Airio (2006) on verrannut eri kielenkäsittelymenetelmien vaikutuksia hakutehokkuuteen yksi- ja kaksikielisessä tiedonhaussa. Mukana vertailtavina kielinä ovat niin suomi, englanti, saksa kuin ruotsikin. Neljä tutkimuksessa vertailtavaa hakemistoa ovat taivutusmuotoinen hakemisto, karsittu hakemisto, perusmuotoistettu hakemisto sekä perusmuotoistettu ja ositettu hakemisto. Airion tutkimuksen tulokset osoittavat, että ruotsin kielessä yksikielisen tiedonhaun osalta perusmuotoistettu ja ositettu hakemisto on tehokkuudeltaan parhain. Pelkästään perusmuotoistettu hakemisto antaa 19,1 prosenttia huonomman tuloksen kuin se perusmuotoistettu hakemisto, jossa yhdyssanat on myös ositettu (Airio 2006: 261). Ahlgrenin (2004) tavoin Airio ei ole osittanut yhdyssanoja hakuvaiheessa, joten molemmat tutkimukset jättävät tämän kysymyksen avoimeksi.

Per Ahlgren onkin täydentänyt vuoden 2004 väitöskirjaansa lisätutkimuksella (Ahlgren & Kekäläinen 2006), jossa yhdyssanojen osittaminen on huomioitu tallennusvaiheen lisäksi myös kyselyissä. Tutkimuksessa vertaillaan seitsemää eri menetelmää, joiden joukossa on neljä perusmuotoistamista hyödyntävää menetelmää. Perusmuotoistamismenetelmissä on hyödynnetty niin yhdyssanojen automaattista kuin valikoivaakin osittamista sekä yhdyssanojen eliminointiperiaat-

teen käyttöä. Näitä menetelmiä verrataan taivutusmuotoiseen hakemistoon ja karsintaan. (Ahlgren & Kekäläinen 2006: 687.) Parhaiten menetelmistä menestyy sananmuotojen katkaiseminen taivutusmuotoisessa hakemistossa. Sitä seuraavien karsinnan ja eri perusmuotoistamismenetelmien välillä ei ole suurta eroa. (Ahlgren & Kekäläinen 2006: 690) Tulosten pohjalta vedetty johtopäätös on se, että ellei tallennusvaiheessa oteta millään tavalla huomioon sananmuotojen morfologista variaatiota, hakuavaimet on katkaistava kyselyvaiheessa (Ahlgren & Kekäläinen 2006: 692).

Ahlgrenin & Kekäläisen (2006) tutkimuksessa on siis täydennetty väitöskirjan puutteita, mutta yksi olennainen menetelmien välinen vertailu puuttuu yhä. Vertailtavissa menetelmissä ei ole mukana sellaista yhdistelmää, jossa hakuavaimina käytettäisiin jakamattomia yhdyssanoja. Tämän vuoksi tutkimuksen perusteella ei voida tehdä päätelmiä siitä, onko hakeminen ositetuilla yhdyssanoilla tehokkaampaa kuin osittamattomilla yhdyssanoilla hakeminen. Tähän kysymykseen pro gradu -tutkielmani pyrkiikin vastaamaan. Lisäksi Ahlgren keskittyy tätä tutkimusta enemmän hakemistojen ja hakuavaimien yhdistelmien vertailuun, siinä missä tässä tutkimuksessa käytetään kahta vakioitua indeksointitapaa (eliminoimaton ja eliminoitu kanta) ja verrataan erilaisia kyselynmuodostamistapoja toisiinsa (lisää luvussa 6.4).

Gunnarsson ja Petersson (2005) ovat tutkineet maisterintutkielmassaan kyselyn laajentamista ruotsin kielessä taivutusvarianttien ja yhdyssanojen osittamisen avulla. Tutkimustuloksien perusteella vedetty johtopäätös on, että yhdyssanojen osittaminen ei kannata, jos 1) yhdyssana on erisnimi, 2) yhdyssanan osat esiintyvät laaja-alaisesti dokumenttikokoelmassa ja 3) yhdyssanan osat eivät yksin esiintyessään ilmaise yhdyssanan merkitystä (Gunnarsson & Petersson 2005: 47). Tutkittavia hakuaiheita tutkimuksessa on 29 kappaletta.

Carlberger ja muut (2001) ovat tutkineet ruotsin kieltä varten kehitetyn karsinta-algoritmin tehokkuutta ruotsinkielisessä tiedonhaussa. Tutkimuksen tulokset osoittavat, että karsinta parantaa ruotsin kielessä sekä saantia että tarkkuutta.

Ruotsin kieli kuuluu germaanisten kielten pohjoisgermaaniseen haaraan, jossa muita kieliä ovat norja, tanska, fääri ja islanti. Länsigermaanisia kieliä puolestaan ovat muun muassa englanti ja saksa, mikä selittää yhteneväiset piirteet ruotsin ja saksan kielen välillä. (Karlsson 2006: 264) Molemmissa kielissä on paljon yhdyssanoja ja yhdyssanojen välissä olevat fogemorfeemit ovat yleisiä. Näin ollen saksan kieltä koskevat tutkimustulokset ovat kiinnostavia myös ruotsin kielen

näkökulmasta. Braschler ja Ripplinger (2004) ovat tutkineet saksan kielen osalta eri kielenkäsittelymenetelmiä kuten karsintaa ja yhdyssanojen osittamista sekä tallennus- että hakuvaiheessa. Tutkimuksessa on tutkittu menetelmiä eripituisilla kyselyillä. Yhdyssanojen osittamisesta todetaan olevan hyötyä saksan kielessä erityisesti lyhyissä kyselyissä. Yhdyssanojen osittamista hyödyntävät menetelmät menestyvät paremmin kuin ne menetelmät, joissa osittamista ei hyödynnetä, kuten esimerkiksi karsintaa hyödyntävät menetelmät (Braschler & Ripplinger 2004: 313).

Myös suomen kielessä on runsaasti yhdyssanoja ja kieltä on tutkittu tiedonhaun näkökulmasta. Riitta Alkula (2000) on tutkinut väitöskirjassaan suomen kielen erityispiirteiden vaikutuksia tiedonhakuun vertailemalla kuutta erilaista hakumenetelmää. Alkula on kiinnittänyt huomiota sanojen perusmuotoistamiseen ja yhdyssanojen osittamiseen, mikä näkyy yhtenä vertailtavana hakumenetelmänä. Väitöskirja on osa FULLTEXT-projektia[1] (Alkula & Honkela 1992), jonka yhtenä tavoitteena on ollut tutkia yhdyssanojen erilaisia tallennus- ja hakuvaihtoehtoja. Suomen kieltä koskevat tulokset ovat suuntaa antavia, mutta suomen kieleen liittyviä tutkimustuloksia ei voi soveltaa ruotsin kieleen kovinkaan onnistuneesti, koska kielet eroavat paljon toisistaan.

Kimmo Kettusen väitöskirja (2007) käsittelee pääasiassa suomen kielen morfologista käsittelyä tiedonhaussa. Mukana on myös ruotsia, saksaa ja venäjää koskevia tutkimustuloksia. Ruotsin kielen osalta aineistona on käytetty CLEF2003-kokoelmaa. Tutkimuksen ensimmäisessä osassa verrataan perusmuotoistamista, taivutusvartaloiden tuottamista ja karsintaa käsittelymenetelminä. Suomen kielen osalta tehokkain menetelmä on perusmuotoistaminen, mutta toisena tulevaan taivutusvartaloiden tuottamiseen ei ole suurta eroa. (Kettunen 2007: 50.) Lisäksi tutkimuksessa tutkitaan taivutusvartaloiden tuottamista vielä tarkemmin ja verrataan pitkiä ja lyhyitä kyselyitä toisiinsa. Tutkimuksessa kartoitetaan FCG-menetelmän (frequent case form generation) käyttöä. Menetelmä toimii siten, että hakuavaimina annetuista substantiiveista ja adjektiiveista käytetään haussa vain niiden tilastollisesti keskeisiä taipuneita muotoja. Ruotsin kielen osalta Kettunen (2007: 54) toteaa, että FCG-menetelmä toimii hyvin sekä pitkissä että lyhyissä kyselyissä. Paras vertailluista menetelmistä on ruotsin kielen osalta kuitenkin perusmuotoistaminen ja yhdyssanojen osittaminen. Kaiken kaikkiaan tutkimuksessa todetaan, että sananmuotoja tuottavat ohjelmat soveltuvat hyvin sananmuotojen käsittelyyn morfologisesti mutkikkaissa kielissä.

[1] FULLTEXT-projekti kantaa myös nimeä Suomenkielisten tekstitietokantojen tallennus- ja hakutekniikat.

Hollink ja muut (2004) ovat vertailleet morfologisen käsittelyn menetelmiä muutamien eurooppalaisten kielten kesken. Mukana vertailussa ovat hollanti, englanti, suomi, ranska, saksa, italia, espanja sekä ruotsi. Vertailtuja menetelmiä ovat karsinta, perusmuotoistaminen, yhdyssanojen osittaminen sekä n-grammit. N-grammeilla tarkoitetaan merkkijonokokonaisuuksien jakamista n-grammeihin eli n-mittaisiin osiin. Digrammit ovat kahden merkin pituisia, trigrammit kolmen merkin pituisia, 4-grammit neljän merkin pituisia ja niin edelleen. Tausta-ajatuksena on merkkijonoketjujen samankaltaisuuden vertaaminen. Hollinkin ja muiden tutkimuksessa yhdyssanojen osittaminen koettiin hyödylliseksi menetelmäksi hollannin, suomen, saksan ja ruotsin kielten osalta. Myös karsinta oli suurimmassa osassa kieliä varteenotettava menetelmä. N-grammeista erityisesti 4-grammien käyttö paransi hakutuloksia.

Hedlundin ja muiden (2001: 148) mukaan skandinaavisia kieliä koskevia tutkimustuloksia voitaisiin soveltaa toinen toisiinsa, koska skandinaavisten kielten piirteet muistuttavat toisiaan. Valitettavasti skandinaavisia kieliä koskevia kansainvälisesti merkittäviä tutkimuksia on kuitenkin tehty suhteellisen vähän.

Pirkola (2001) esittää artikkelissaan morfologista typologista luokittelua eri kielten välille sen mukaan, millaista sananmuotojen käsittelyä kielet vaativat. Perinteisesti kielitieteessä on käytetty typologisen vertailun välineenä synteesi-indeksiä (index of synthesis) ja fuusioindeksiä (index of fusion). Edellinen viittaa taivutuksen määrään kielessä. Jälkimmäisellä viitataan puolestaan siihen, kuinka hyvin sanan sisältämät morfeemit ovat erotettavissa toisista morfeemeista. (Pirkola 2001: 336.) Pirkola soveltaa ideaa edelleen tiedonhakuun ja on kehitellyt erilaisia vertailuindeksejä, joiden avulla voidaan päätellä, millainen morfologinen käsittely kunkin kielen osalta on hyödyllistä. Taivutusindeksi IIS (inflectional index of synthesis) muodostetaan laskemalla taivutusmorfeemien määrä ja jakamalla se kaikkien tekstiotoksen sisältämien sanojen määrällä. Johtamisindeksi DIS (derivational index of synthesis) muodostetaan laskemalla johtamismorfeemien määrä ja jakamalla se koko sanamäärällä. Yhdyssanaindeksi CIS (compoud index of synthesis) muodostetaan laskemalla yhdyssanamorfeemien määrä ja jakamalla se kokonaissanamäärällä. (Pirkola 2001: 337)

Pirkolan ajatuksena on, että kun nämä indeksit lasketaan eri kielten osalta, voidaan helpommin vertailla kieliä ja tehdä päätelmiä siitä, millaiset morfologisen käsittelyn menetelmät sopivat parhaiten kullekin kielelle. Tässä tutkimuksessa tutkitaan yhdyssanojen lukumäärää, joten viimeksi mainittu indeksi on mahdollista toteuttaa. Fuusioindeksistä Pirkola on kehittänyt kolme morfolo-

gisen fuusion indeksiä, joista yksi on yhdyssanoihin liittyvä MorphCIF, joka muodostetaan laskemalla fuusioituneiden yhdyssanojen määrä ja jakamalla se koko sanamäärällä. Samanlaiset indeksit on mahdollista laskea myös fuusioituneiden taivutusmuotojen ja johdosten osalta. (Pirkola 2001: 338) Fuusioitunut yhdyssana vastaa tässä siis ei-kompositionaalista yhdyssanaa. Fuusion lopputuote on jotakin muuta kuin sen osien summa antaisi olettaa. Ajatuksena on se, että mitä matalampi kukin indeksi on, sitä vähemmän kielessä on ongelmia kyseisen ilmiön osalta.

Myös semantiikan tasolla on mahdollista suorittaa kielten typologista vertailua. Pirkolan mukaan yhdyssanat ja sanojen johtaminen liittyvät usein merkitysten muutoksiin. Tämän vuoksi myös semantiikan tasolla voidaan laskea indeksejä. Semanttinen fuusioindeksi yhdyssanojen osalta (SemCIF) muodostetaan laskemalla fuusioyhdyssanojen määrä ja jakamalla se kaikkien yhdyssanojen määrällä. Semanttinen fuusioindeksi johtamisen osalta (SemDIF) muodostetaan laskemalla fuusiojohdosten määrä ja jakamalla se kaikkien johdosten määrällä. (Pirkola 2001: 338) Esimerkiksi matala SemCIF viittaa siihen, että suurin osa yhdyssanoista on kompositionaalisia ja yhdyssanojen osiin jakamisesta voisi olla hyötyä tiedonhaussa.

5 TUTKIMUSONGELMA

Pro gradu -tutkielmani yhdistää kielitieteen ja tiedonhaun tutkimusta toisiinsa ja kiinnostuksen kohteena on ruotsin kieli ja kielen sisältämät yhdyssanat. Tarkoituksena on tutkia sekä yhdyssanojen yleisyyttä että sitä, miten ne tulisi ottaa huomioon tiedonhaussa. Tutkimusongelma jakautuukin kahteen osaan:

- Yhdyssanojen määrä ja tyypit ruotsinkielisissä hakuaiheissa ja dokumenteissa
- Yhdyssanojen osittamisen ja yhdysosien eliminoimisen vaikutus hakutuloksiin ruotsin kielessä

Tutkimusongelman ensimmäisessä kohdassa on tarkoituksena selvittää, kuinka paljon ruotsinkielisissä dokumenteissa ja ruotsinkielisissä hakuaiheissa esiintyy yhdyssanoja. Hakuaiheista analyysin kohteena ovat descriptor-kentässä esiintyvät yhdyssanat, joiden määrää verrataan sanojen kokonaismäärään. Hakuaihe jakautuu siis eri kenttiin, joista yksi on descriptor-kenttä (ks. liite 1).

Tutkittava kenttä on sama kenttä, josta kyselyt muodostetaan. Dokumenttien osalta poimitaan otos, josta yhdyssanojen määriä kartoitetaan. Lisää käytännön toteutuksesta on luvassa luvussa 6.

Lisäksi on tarkoitus tutkia eri yhdyssanatyyppien määrää hakuaiheissa ja dokumenttiotoksessa Pirkolan (2001) idean mukaisesti. Pirkola (2001: 343) esittää, että olisi tarpeen tutkia kompositionaalisten eli merkitykseltään läpinäkyvien yhdyssanojen määrää ruotsin kielessä. Näin voidaan tehdä myös päätelmiä siitä, miten ruotsin kieltä kannattaa käsitellä tiedonhaussa. Jos suurin osa yhdyssanoista on kompositionaalisia, voidaan olettaa, että yhdyssanojen osittaminen kannattaa ruotsinkielisessä tiedonhaussa. Kompositionaalisuuden lisäksi kartoitetaan yhdyssanojen sanaluokkajakaumia hakuaiheiden osalta.

Tutkimusongelman toisessa kohdassa tarkoituksena on täydentää Ahlgrenin (2004) sekä Ahlgrenin ja Kekäläisen (2006) tutkimuksia ja tarkastella, kannattaako yhdyssanojen osittaminen kyselyvaiheessa ruotsin kielessä. Vertailtavia menetelmiä ovat kolme erilaista kyselysarjaa, joilla haetaan kahdessa erilaisessa hakemistossa, yhdyssanoiltaan eliminoidussa hakemistossa ja yhdyssanoiltaan eliminoimattomassa hakemistossa. Kolme vertailtavaa kyselynmuodostusmenetelmää ovat kyselyjen muodostaminen a) yhdyssanat perusmuotoistettuina, mutta osittamattomina ja b) yhdyssanat perusmuotoistettuina, ositettuina ja eliminoimattomina sekä c) yhdyssanat perusmuotoistettuina, ositettuina ja eliminoituina. Kyselysarjoista haetaan sekä rakenteisilla että rakenteettomilla kyselyillä. Rakenteisissa kyselyissä hyödynnetään #combine-operaattorin lisäksi synonyymioperaattoria ja läheisyysoperaattoria, kun taas rakenteettomat kyselyt muodostuvat #combine-operaattorista ja sanalistasta (lisää kyselyistä luvussa 6.4.2 ja liitteessä 2).

Tämä tutkimusongelman jälkimmäinen osuus on perinteinen tiedonhaun laboratoriotutkimus, jolle on ominaista se, ettei käyttäjä osallistu tutkimukseen, vaan tiedonhakuprosessi on pitkälle automatisoitu ja hakuaiheet edustavat käyttäjiä. Laboratoriotutkimuksessa keskitytäänkin dokumenttien esityksiin, kyselyihin ja niiden täsmäyttämiseen toisiinsa (Ingwersen & Järvelin 2005: 114–115). Laboratoriotutkimusta on esitelty tarkemmin luvussa 2.6.

Tutkimusongelmaan liittyvät tutkimuskysymykset ovat siis seuraavat:

- Kuinka paljon ruotsinkielisessä tekstissä (hakuaiheiden descriptor-kentässä ja dokumenttiotoksessa) on yhdyssanoja?
- Kuinka suuri osa yhdyssanoista on kompositionaalisia eli merkitykseltään läpinäkyviä yhdyssanoja? Mitä sanaluokkaa hakuaiheissa esiintyvät yhdyssanat edustavat?

- Onko ruotsin kielessä tuloksellista hakea yhdyssanat osittamattomina vai yhdyssanat ositettuina? Miten yhdyssanojen yhdysosien eliminoiminen vaikuttaa hakutuloksiin? Vaikuttaako kyselyn rakenteen muuttuminen hakutuloksiin?

6 AINEISTO JA MENETELMÄT

Tässä luvussa esitellään tutkimuksessa käytettävät aineistot ja analyysimenetelmät.

6.1 Testikokoelmat

Toinen tutkimuksessa käytettävä aineisto on CLEF-kokoelma (Cross Language Evaluation Forum [www] 2008). CLEF-kokoelma on monikielinen kokoelma, joka sisältää myös ruotsinkielisiä dokumentteja ja hakuaiheita sekä niihin liittyvät relevanssiarviot. Monikielisyytensä vuoksi kokoelmaa voidaan käyttää aineistona myös kieltenvälistä tiedonhakua koskevassa tutkimuksessa. Tutkimuksessa käytettävä ruotsinkielinen dokumenttikokoelma sisältää 142 819 uutisdokumenttia, jotka ovat Ruotsin tietotoimiston TT:n (Tidningarnas telegrambyrå) uutisia vuosilta 1994−1995. Käytettävät hakuaiheet ja relevanssiarviot ovat vuosilta 2002−2003. Liitteessä 1 on esimerkit CLEF2003- ja CLEF2002-hakuaiheista, joiden descriptor-kentistä kyselyt muodostetaan.

Tutkimuksessa käytetään rinnakkain kolmea eri aineistoa. CLEF2002- ja 2003-aineistojen lisäksi aineistona käytetään Per Ahlgrenin kokoelmaa. Kokoelma koostuu 51 kysymyksestä, jotka ovat vuoden 2000-2001 CLEF-hakuaiheista muokattuja. Liitteessä 1 on esimerkki yhdestä hakuaiheesta. Dokumenttikokoelma koostuu Helsingborgs Dagbladin ja Göteborgs Postenin uutisartikkeleista, joita on kaikkiaan 161 336 kappaletta. Aineiston etuja on muun muassa se, että siinä on käytetty moniportaista relevanssiarviointia. Toisaalta Ahlgrenin aineiston ongelmana on relevanssikorpusten muodostustapa. Keskimäärin on arvioitu 100 dokumenttia aihetta kohden, ja joukossa on paljon dokumentteja, joita ei ole arvioitu. Kolmea eri aineistoa käyttämällä voidaan vertailla niillä saatuja tuloksia ja näin saadaan myös luotettavampia tuloksia.

6.2 Yhdyssanojen määrän ja tyyppien analysoiminen hakuaiheissa

Ensiksikin yhdyssanojen määrää ja tyyppejä kartoitetaan CLEF- ja Ahlgren-aineistojen hakuaiheissa. Hakuaiheista analyysin kohteena ovat descriptor-kentässä esiintyvät yhdyssanat, joiden määrää verrataan sanojen kokonaismäärään. Tutkittava kenttä on siis sama kenttä, josta kyselyt muodostetaan.

Yhdyssanojen tunnistaminen perustuu määritelmään, jonka mukaan yhdyssana koostuu yhdestä tai useammasta osasta, joista kukin voi esiintyä itsenäisenä sanana. Ongelmatapauksissa lähteenä käytetään Svenska Akademiens Ordlistaa (2006), joka kattaa nykyruotsin yleisimmät sanat. Myös Internetistä löytyvää historiallista sanakirjaa Svenska Akademiens ordbokia (2008) käytetään lähteenä ongelmatapauksissa. Liljestrandin (1993: 33) mukaan sanakirjat eivät sisällä suurinta osaa yhdyssanoista, ja tämä vaikeuttaa analyysia. Sanakirjoissa esiintyvät ainoastaan sellaiset yhdyssanat, jotka ovat yleisesti käytössä ja joiden merkitys on usein leksikaalistunut. Analyysin taustalla on kuitenkin oletus, jonka mukaan tilapäiset yhdyssanat ovat helposti tunnistettavissa ilman sanakirjan apua. Myös SWETWOL on avuksi ongelmallisissa tapauksissa, koska on mahdollista tarkistaa, miten SWETWOL tulkitsee sanan. Kun yhdyssanoja lasketaan, otetaan huomioon vain sanat, ei yhdyssanan sisältämiä toisia yhdyssanoja. Esimerkiksi moniosainen yhdyssana **järn-väg-station** lasketaan yhdeksi yhdyssanaksi, vaikka siihen sisältyykin myös toinen yhdyssana.

6.3 Otoksen poimiminen aineistosta

Hakuaiheiden analysoimisen jälkeen yhdyssanojen määrä ja tyypit käydään läpi dokumenttien osalta. Tutkimuksessa käytettävä aineisto sisältää tuhansia dokumentteja, joten yhdyssana-analyysissa käytetään 200 dokumentin otosta. Seuraavaksi esittelen otoksen poimimiseen liittyviä käsitteitä.

Perusjoukolla tarkoitetaan tutkimuksen kohdejoukkoa, tässä tutkimuksessa perusjoukkona toimivat ruotsinkielisen sanomalehtiaineiston dokumentit. *Otos* puolestaan tarkoittaa otantamenetelmän avulla perusjoukosta poimittua havaintoyksiköiden joukkoa. (Vilkka 2007: 51.) Otoksen poimimisessa korostetaan otoksen *edustavuuden* merkitystä, millä tarkoitetaan sitä, että valitussa otoksessa tulee olla samanlaiset ominaisuudet samassa suhteessa kuin perusjoukossa (Holopainen & Pulkkinen 1995: 21). Myös Vilkka (2007: 56−57) korostaa, että otos on kokonaiskuva

perusjoukosta, minkä vuoksi jokaisella perusjoukon havaintoyksiköllä tulisi olla yhtäläiset mahdollisuudet valikoitua otokseen. Otos ei Vilkan mukaan kuitenkaan koskaan täysin kuvaa perusjoukkoa.

Yksinkertaisessa satunnaisotannassa (simple random sampling) otoksen poimiminen suoritetaan arpomalla siten, että jokaisella havaintoyksiköllä on yhtä suuri todennäköisyys tulla valituksi. Mikäli perusjoukko jakautuu heterogeenisiin ryhmiin tai mikäli ryhmät ovat homogeenisiä ja eroteltavissa, voidaan käyttää *ositettua otantaa* (stratified random sampling). Kutakin ryhmää kutsutaan *ositteeksi* ja kustakin ositteesta poimitaan erikseen satunnaisotos. *Tasaisessa kiintiöinnissä* jokaisesta ositteesta poimitaan yhtä monta havaintoyksikköä. *Suhteellisessa kiintiöinnissä* otantasuhde on jokaisesta ositteesta sama. (Holopainen & Pulkkinen 1995: 22) Havaintoyksikköjä poimitaan siis prosentuaalisesti sama määrä kustakin ositteesta. Kolmas tapa on poimia kiintiöimällä optimaalisesti eli tässä tavassa huomioidaan ositteen koko, hajonta sekä kustannukset. (Vilkka 2007: 54−55) Pickard (2007: 62) huomauttaa, että kunkin ryhmän tulisi olla edustettuna otoksessa yhtä suuressa määrin ottaen huomioon ryhmän koko suhteessa koko perusjoukkoon. Myös Vilkan (2007: 55) mukaan ositteissa tulisi olla edustettuina perusjoukon samanlaiset ominaisuudet.

Systemaattista otantaa puolestaan käytetään niissä tapauksissa, kun perusjoukkoa ei voida tarkkaan määrittää. Systemaattisessa otannassa lasketaan ensin suhdeluku N/n (N=perusjoukko, n=otos), jonka avullaan saadaan poimintaväli. Ensimmäinen poimittava havaintoyksikkö on poimintavälin mukainen joka n:s havaintoyksikkö. Tämän jälkeen seuraavat yksiköt poimitaan tasaisin välein järjestyksessä joka n:s tilastoyksikkö. (Holopainen & Pulkkinen 1995: 23)

6.3.1 Otoksen poiminnan toteutus

Yhdyssanojen määrää dokumenteissa tutkitaan aineiston koon vuoksi otoksen avulla. Selvyyden vuoksi aineisto on jaettu kahtia CLEF-aineistoon (sisältää vuoden 2002 ja 2003 aineistot) ja Per Ahlgrenin kokoelmaan. Näistä edellinen sisältää yhteensä 142 819 dokumenttia ja jälkimmäinen 161 336 dokumenttia. Otoksen koko on 200 dokumenttia. Puolet 200 dokumentista poimitaan CLEF-aineistosta ja puolet Ahlgrenin kokoelmasta.

CLEF-aineisto (CLEF2003 ja CLEF2002) jakautuu 24 tiedostoon, joten poiminnassa hyödynnettiin ositettua otantaa siten, että poimittiin satunnaisesti arpomalla tasaisesti sama määrä doku-

mentteja joka tiedostosta (tasainen kiintiöinti). 100 jaettuna 24:lla on 4,166, joten tasaisen kiintiöinnin määräksi valikoitui neljä dokumenttia tiedostoa kohti. Näin kokonaismääräksi tuli 96 dokumenttia. Jäljelle jäävät 104 dokumenttia poimittiin Per Ahlgrenin aineistosta, joka jakautuu kahteen eri tiedostoon. Kummastakin tiedostosta poimittiin siis 52 dokumenttia satunnaisesti arpomalla.

Käytännön tasolla dokumenttien tunnisteet paikannettiin Unix-järjestelmässä grep-komennon avulla poimimalla dokumenttitunnisteen sisältävät rivit, jolloin rivejä muodostui yhtä monta kuin dokumentteja. Esimerkiksi CLEF-aineistosta poimittiin jokaisesta tiedostosta rivit, joissa on <DOCID>. Tämän jälkeen rivit siirrettiin Excel-taulukkolaskentaohjelmaan, jossa itse dokumenttien poiminta tapahtui. Excelin rivinumero vastaa dokumentin järjestysnumeroa, mikä helpotti tietyllä rivillä olevan dokumenttinumeron poimimista. Ahlgrenin kokoelman Göteborgs Posten -aineistossa järjestysnumero on sama kuin dokumentin numero, koska dokumenttien numerointi alkaa alusta numerosta yksi ja jatkuu numerojärjestyksessä. Ahlgrenin kokoelman Helsingborgs Dagblad-aineisto oli jaettu kahtia (dokumenttien numerot eivät olleet järjestyksessä), joten siinä hyödynnettiin samaa menetelmää kuin CLEF-aineiston kanssa.

6.3.2 Yhdyssanojen laskeminen otoksesta

Dokumenttianalyysin aineistona käytetään siis 200 dokumentin aineistoa. Aineistosta poimittiin SWETWOL:n avulla sanat, jotka SWETWOL tulkitsee yhdyssanaksi ja sanat, joita SWETWOL ei tunnista yhdyssanaksi. Määritelmässä yhdyssanaksi lasketaan sana, joka jakautuu vähintään kahteen itsenäiseen osaan. SWETWOL:n tekemät yhdyssanatulkinnat tunnistaa siitä, että sanojen välissä on joko merkki # tai |. SWETWOL voi olla monitulkintainen, joten kutakin SWETWOL:n luentaa kohden laskettiin vain yksi merkki, joko # tai |. Muutoin yhdyssanojen määrä kasvaa helposti, koska monista sanoista on useita erilaisia luentoja.

Lisäksi otettiin huomioon sanojen kirjoitusasu, esimerkiksi se, kirjoitetaanko sana pienellä vai isolla ja onko sanojen välissä muita välimerkkejä kuten yhdysviivoja. Näillä päätöksillä on vaikutusta SWETWOL:n tekemiin tulkintoihin. SWETWOL:n käyttöön liittyikin seuraavanlaisia huomioita:

- Maan nimet pitää kirjoittaa pienellä, jotta SWETWOL tunnistaa, esimerkiksi **skottland**

- Myös erisnimen sisältävät yhdyssanat pitää kirjoittaa pienellä, jotta SWETWOL tunnistaa yhdyssanaksi, esimerkiksi **balkankonflikten**
- Sama koskee osaa yhdysviivallisista sanoista, esimerkiksi SWETWOL tunnistaa sanan **cd-brännare**, mutta ei **CD-brännare**

Sekä CLEF-aineistosta että Per Ahlgren-aineistosta poimittiin siis eri tiedostoihin yhdyssanat ja ne sanat, joita SWETWOL ei tunnista yhdyssanoiksi. Näiden tiedostojen avulla voitiin selvittää sekä yhdyssanojen määrää että tyyppejä dokumenttiotoksessa.

6.4 Eräajojen toteuttaminen

Yhdyssanojen määrän ja tyyppien kartoittamisen lisäksi tässä tutkimuksessa tutkitaan erilaisten kyselynmuodostustapojen vaikutusta hakutuloksiin ruotsin kielessä. Tämä osa tutkimuksesta on tiedonhaun laboratoriotutkimus. Seuraavaksi esitellään tutkimuksen koeasetelmaa.

6.4.1 Tiedonhakujärjestelmä ja hakemistot

Tiedonhakujärjestelmänä käytetään Indri-järjestelmää, joka perustuu kielimallien ja päättelyverkon käyttöön (Strohman 2004: 2). Indri on Massachusettsin ja Carnegie Mellonin yliopistojen yhteistyönä tutkimuskäyttöön tehty, vapaasti jaeltu ohjelmisto[2]. Indri mahdollistaa morfologisen analyysiohjelman käyttämisen tallennuksen yhteydessä. Tässä tutkielmassa Indri-järjestelmässä käytettävän tietokannan hakemisto on rakennettu morfologisen analyysiohjelman SWETWOL:n tulosteista. Kyselyjen täsmäytyksessä käytetään kielimallia Dirichlet-tasoituksella. Tulosjoukon koko on 1000 dokumenttia kyselyä kohden.

Kyselyjen syntaksissa on useita erilaisia vaihtoehtoja. #combine-operaattorin tehtävänä on yhdistää hakuavaimet. Operaattori laskee kyselyn faseteille keskiarvon. #uwn-operaattori (unordered window) on läheisyysoperaattori, jossa n määrittelee sanojen etäisyyden toisistaan. #uwn-operaattorilla erotetut sanat voivat esiintyä missä tahansa järjestyksessä mutta tietyllä etäisyydellä toisistaan. Lisäksi on mahdollista hyödyntää esimerkiksi #syn-rakennetta. #syn-operaattorin sulkujen sisällä olevia sanoja kohdellaan saman sanan eri esiintyminä eikä niille lasketa erikseen painoja. (Strohman 2004: 2.)

[2] Lisätietoja http://www.lemurproject.org/indri/

CLEF-kokoelma ja Ahlgrenin kokoelma on indeksoitu kahteen erilaiseen hakemistoon. Toinen on perusmuotoistettu ja yhdyssanoiltaan ositettu, joskin yhdyssanojen eliminoimista ei ole hyödynnetty. Toinen on perusmuotoistettu ja yhdyssanojen osalta sekä ositettu että eliminoitu. Hakemistoihin viitataan tässä tutkielmassa myöhemmin termeillä eliminoitu kanta ja eliminoimaton kanta.

6.4.2 Kyselysarjat

Vertailtavat menetelmät muodostavat kolme erilaista kyselysarjaa, joilla tehdään hakuja kahteen eri hakemistoon. Yhteensä vertailussa on siis kuusi erilaista yhdistelmää. Eliminoidussa ja eliminoimattomassa kannassa verrataan kummassakin kyselyjen muodostamista a) yhdyssanat perusmuotoistettuina ja osittamattomina, b) yhdyssanat perusmuotoistettuina, ositettuina ja eliminoimattomina sekä c) yhdyssanat perusmuotoistettuina, ositettuina ja eliminoituina.

Tutkimuksessa käytetään siis kolmen eri aineiston hakuaiheita, joista on muodostettu kolme erilaista kyselysarjaa. CLEF2003-aineistossa hakuaiheita on 54 kappaletta. CLEF2002-aineisto sisältää 49 hakuaihetta. Per Ahlgrenin aineistossa hakuaiheita on puolestaan 51 kappaletta. Tutkimuksessa käytetään kahdella eri tavalla muodostettuja kyselyitä, rakenteettomia ja rakenteisia kyselyitä.

Molempien kyselyversioiden toteuttamisessa käytettiin apuna Query Converter-ohjelmaa, jonka avulla eri kyselytiedostot muokattiin hakuaiheiden descriptor-kentistä. Liitteessä 1 on esimerkit kunkin aineiston yhdestä hakuaiheesta. Liitteessä 2 puolestaan on esitetty esimerkit kunkin aineiston eri kyselytyypeistä. Query Converter teki valmiiksi kyselyjen perussyntaksin. Query Converterin avulla oli myös mahdollista määritellä eri käsittelyjen järjestys, esimerkiksi se, missä vaiheessa sulkusanat poistetaan kyselyistä. Tässä tapauksessa sulkusanat poistettiin sekä ennen että jälkeen perusmuotoistamisen. Rakenteettomissa kyselyissä kyselyjen syntaksi koostuu pelkästä #combine-operaattorista. Kyselyjä voidaan kutsua myös bag of words -kyselyiksi, koska kyselyt ovat #combine-operaattorilla yhdistettyjä sanalistoja. Rakenteisissa kyselyissä syntaksi koostuu #combine- ja #syn-operaattoreista. Rakenteisissa kyselyissä #syn-operaattoria käytetään erottamaan saman sananmuodon eri variantit, esimerkiksi **nederland**, **nederländ**. Rakenteettomissa kyselyissä yhdyssanat ja niiden osat esiintyvät peräkkäin samassa sanalistassa. Rakentei-

sissa kyselyissä yhdyssanojen osittamisen myötä syntyneitä yhdysosia käsiteltiin lisäämällä kyselyihin eri yhdysosat yhdistävä #uw20-läheisyysoperaattori.

#uw20-läheisyysoperaattori edellyttää, että yhdyssanojen etu- ja loppuosat esiintyvät vähintään 20 sanan etäisyydellä toisistaan. Läheisyysoperaattorin käyttö mahdollistaa sen, että yhdyssanojen osia ei oteta huomioon mistä tahansa tekstin osasta, vaan yhdyssanojen osien käsittelemiä asioita esiintyy 20 sanan etäisyydellä. Läheisyysoperaattorin käyttö siis vähentää väärien yhdistelmien määrää. Toisaalta läheisyysoperaattorin käytössä on myös haittapuolia, sillä se rajaa hakua. Yhdyssanojen osittamista on pidetty hyödyllisenä käsittelyvaihtoehtona sen vuoksi, että se tuo esiin yhdyssanojen loppuosat hakuavaimina. Esimerkiksi tekstissä saatetaan mainita ensimmäisen kerran koko yhdyssana, johon viitataan myöhemmin pelkällä loppuosalla. Tällöin läheisyysoperaattori kuitenkin edellyttää myös yhdyssanan etuosan esiintymistä 20 sanan etäisyydellä. Yksi haittapuoli liittyy myös tutkimustulosten tulkitsemiseen. Tarkasteltaessa hakutuloksia ei ole aivan selvää, mikä osa hakutuloksesta on yhdyssanojen osittamisen vaikutusta, mikä puolestaan läheisyysoperaattorin ansiota. Tämän vuoksi tutkimuksessa toteutettiin eräajot myös rakenteettomilla kyselyillä (bag of words), joissa kyselyn sanoja yhdistää ainoastaan #combine-operaattori.

Eräajojen tulosten pohjalta tehtiin vielä tilastolliset testit, jotta pystyttäisiin tekemään päätelmiä siitä, onko vertailtavien menetelmien välillä tilastollisesti merkitseviä eroja. Smucker, Allan ja Carterette (2007) suosittelevat parametristen testien käyttöä. Tutkimuksen koeasetelmaan sopiva testi on yksisuuntainen varianssianalyysi.

6.4.3 Relevanssikorpukset

CLEF-aineistojen osalta relevanssi on arvioitu binäärisesti siten, että dokumentti on joko relevantti tai epärelevantti. Per Ahlgrenin kokoelmassa relevanssiarviot on tehty dokumenteille neliportaista (0−3) asteikkoa käyttäen. Relevanssiarvioissa arvo 0 on annettu dokumenteille, jotka ovat täysin epärelevantteja. Dokumentit eivät sisällä mitään haun aiheeseen liittyvää tietoa. Arvon 1 saavat puolestaan dokumentit, jotka ovat marginaalisesti relevantteja. Näissä dokumenteissa kyllä viitataan haun aiheeseen, mutta ne eivät sisällä mitään muunlaista tietoa aiheesta. Arvioinnissa arvon 2 ovat saaneet dokumentit, jotka ovat jossain määrin relevantteja. Dokumentit sisältävät jo enemmän tietoa haun aiheesta, mutta kaikkia eri näkökulmia ei ole käsitelty. Arvon

3 ovat puolestaan saaneet dokumentit, jotka ovat erittäin relevantteja ja jotka käsittelevät pääasiassa vain hakuaiheessa mainittua aihetta useista eri näkökulmista. (Ahlgren 2004: 85.)

Käytännön eräajoissa relevanssiarviot jakautuvat Per Ahlgrenin kokoelmassa kolmeen eri relevanssitasoon: liberaaliin, normaaliin ja tiukkaan. Liberaali relevanssitaso laskee relevanteiksi dokumenteiksi arvon 1, 2 tai 3 saaneet dokumentit. Normaalilla relevanssitasolla relevantteja ovat ainoastaan arvon 2 tai 3 saaneet dokumentit. Tiukalla relevanssitasolla relevanteiksi lasketaan vain arvon 3 saaneet dokumentit ja kaikki muut dokumentit tulkitaan epärelevanteiksi.

7 YHDYSSANOJEN MÄÄRÄ JA TYYPIT HAKUAIHEISSA JA DOKUMENTEISSA

Tässä luvussa tarkastelen yhdyssanojen määrää ja eri yhdyssanatyyppien jakaumia ruotsinkielisissä hakuaiheissa ja dokumenteissa. Mukana vertailussa on kolme eri aineistoa: CLEF2003, CLEF2002 sekä Per Ahlgrenin kokoelma.

7.1 Yhdyssanojen ongelmallinen luonne

Tässä tutkimuksessa tehty yhdyssanojen analyysi paljasti yhdyssanojen yllättävän monitulkintaisen ja ongelmallisen luonteen. Tutkielmassa käytetty määritelmä on se, että yhdyssana koostuu kahdesta tai useammasta osasta ja että nämä osat voivat toimia itsenäisinä sanoina. Selvittäessäni tarkemmin, mitä yhdyssanalla tarkoitetaan eri lähteissä, törmäsin kuitenkin ongelmiin. Svenska Akademiens ordlista (2006) sisältää nykyruotsin keskeisimmät sanat ja sanakirjaan on myös merkitty johdokset ja yhdyssanat. Kävin tutkimusaineiston hakuaiheiden sisältämät yhdyssanat läpi SAOL-sanakirjaa apuna käyttäen ja huomasin, että SAOL on tulkinnoissaan liberaali ja hyväksyy sanoja yhdyssanoiksi vaihtelevin perustein. Sana voidaan tulkita yhdyssanaksi muun muassa kielihistoriallisin perustein. Esimerkiksi sana **irländsk** jakautuu osiin **ir-ländsk**. Aikaisemmin **ir** on ollut sana, ja kielihistorialliset syyt siis perustelevat sanan tulkintaa yhdyssanaksi.

Lisäksi SAOL tulkitsee yhdyssanoiksi sanoja, jotka tässä tutkielmassa käytetyn määritelmän mukaan ovat pikemminkin johdoksia. Esimerkiksi **olaglig** jakautuu osiin **o-laglig**, eikä prefiksi **o** voi toimia itsenäisenä sanana. Toinen esimerkki tällaisesta sanasta on **olycka**. Kolmanneksi

SAOL tulkitsee yhdyssanoiksi yhdyssanaverbit ja niistä johdetut substantiivit. Muun muassa verbistä **avskeda** on muodostettu substantiivi **avsked**, joka on SAOL:n mukaan yhdyssana. Substantiivin yhdysosista ei kuitenkaan voi enää päätellä sen kokonaismerkitystä., vaan kokonaismerkityksestä tulee mielivaltainen **av** '-sta, -stä, -lta, -ltä' ja **sked** 'lusikka'. SAOL tulkitsee myös yhdyssanoiksi partikkeliverbeistä muodostetut substantiivit kuten **inslagning**, **uttalande**, **påföljd**, **antal**, **förslag**, **förekomst**, **påtryckning**, **avfall**, **uppmaning**, **upptäckt**, **införande**, **avtal**, **utvecklare**, **användning**, **intag**, **företag** ja **upphov**. SAOL tulkitsee yhdyssanoiksi myös lähes kaikki yhdyssanaverbit, joita SWETWOL ei lue yhdyssanoiksi: **påverka**, **avse**, **avbryta**, **utföra**, **utgöra**, **inbegripa**, **utropa**, **avgå**, **pågå** ja **avsätta**. Näiden lisäksi SAOL-tulkintojen joukossa on muita sanoja, jotka eivät täytä tässä tutkimuksessa käytettävää yhdyssanamääritelmää: **förorening**, **ihjäl**, **åtgärd**, **område**, **allmänhet**, **inhemsk**, **giftermål**, **utrikes**, **samhälle**, **dessutom** ja **varför**.

Toisena lähteenä yhdyssana-analyysissa olen käyttänyt morfologista analyysiohjelmaa SWETWOL:ia. Ohjelman tekemä analyysi ei myöskään ole täysin ongelmatonta. SWETWOL tunnistaa yhdyssanoiksi sanoja, jotka kieliopillisesti ajateltuina eivät ole yhdyssanoja. Esimerkiksi sana **styrkor** (taivutusmuoto sanasta **styrka**) on SWETWOL:n mukaan yhdyssana, joka jakautuu osiin **styr-kor**. Verbi **beskriva** on SWETWOL:n mukaan yhdyssana, joka jakautuu osiin **besk-riva**. Molempiin käyttämiini lähteisiin liittyy siis ongelmia, mikä toisaalta todistaa myös sitä, ettei yhdyssanoista ole olemassa sataprosenttista ja johdonmukaista määritelmää.

7.2 Yhdyssanojen määrä ja tyypit hakuaiheissa

Yhdyssanojen määrää ja yhdyssanatyyppejä tarkastellaan tässä tutkielmassa kolmen eri aineiston osalta. Edellä mainittujen tulkintaongelmien vuoksi SAOL:n tukemia tulkintoja ja SWETWOL:n tukemia tulkintoja käsitellään erikseen sekä verrataan toisiinsa.

7.2.1 Yhdyssanojen määrä

Taulukossa 1 on esitetty yhdyssanojen määrät aineistoissa SAOL-sanakirjan tulkinnan mukaan. Yhdyssanojen määrä vaihtelee 103 ja 127 välillä. CLEF2002 poikkeaa kahdesta muusta vertailtavasta aineistosta pienemmän yhdyssanamäärän perusteella. Keskimäärin koko aineistossa 16,3 prosenttia sanoista on yhdyssanoja. Hedlundin (2002: luku 3) tekemän testin mukaan yhdyssanojen yleisyys 100 000 sanan sanomalehtiaineistossa on suomen kielessä 8,7 prosenttia, ruotsin

kielessä 9,8 prosenttia ja saksan kielessä 10,2 prosenttia. SAOL-tulkintojen mukaan analysoitaessa yhdyssanojen määrä on siis Hedlundin tuloksia suurempi.

TAULUKKO 1. SAOL-yhdyssanat hakuaiheissa

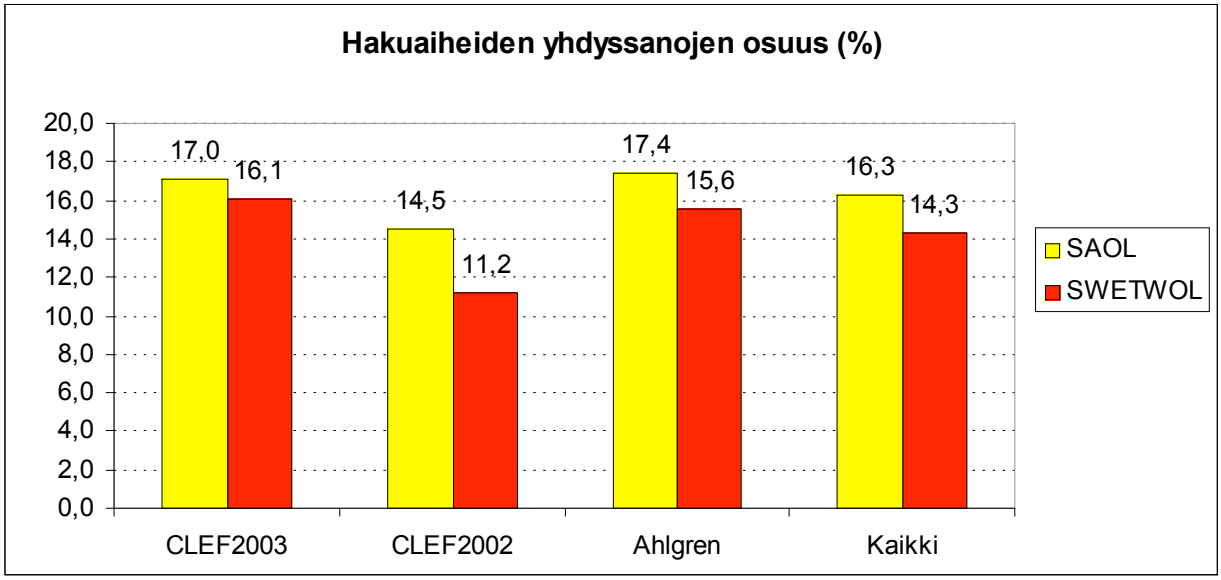
Aineisto	Sanamäärä	SAOL	Prosenttia
CLEF2003	745	127	17,0
CLEF2002	708	103	14,5
Ahlgren	668	116	17,4
Kaikki	2121	346	16,3

Toinen laskentatapa perustuu SWETWOL:n tunnistamiin yhdyssanoihin. SWETWOL-yhdyssanojen määrä vaihtelee aineistossa 79 ja 120 välillä. Myös SWETWOL:n kohdalla CLEF2002-aineisto poikkeaa kahdesta muusta aineistosta pienemmän yhdyssanamääränsä vuoksi. Keskimäärin koko aineistossa 14,3 prosenttia sanoista on yhdyssanoja. Kuten taulukosta 2 nähdään SWETWOL tekee kuitenkin myös paljon virhetulkintoja. Virhetulkintojen osuus SWETWOL-yhdyssanoista on jopa 25 prosentin luokkaa. Virhetulkinnoilla tässä tarkoitetaan vääriä yhdyssanatulkintoja, joita esiteltiin luvussa 7.1.

TAULUKKO 2. SWETWOL-yhdyssanat hakuaiheissa

Aineisto	Sanamäärä	SWETWOL	Prosenttia	Virhetulkinnat	Prosenttia
CLEF2003	745	120	16,1	33	27,5
CLEF2002	708	79	11,2	19	24,1
Ahlgren	668	104	15,6	24	23,1
Kaikki	2121	303	14,3	76	25,1

Hedlund teki analyysinsä käyttäen apunaan SWETWOL:ia, joten siinä mielessä hänen lukunsa ja taulukossa 2 esitetyt luvut ovat vertailukelpoisia. Myös SWETWOL:n osalta yhdyssanojen määrä on aineistossa suurempi kuin Hedlundin tekemässä tutkimuksessa, mikä toisaalta myös korostaa yhdyssanojen roolia tiedonhaussa. CLEF2002, jossa yhdyssanamäärät ovat muita aineistoja pienempiä, on lähimpänä Hedlundin esittämiä lukuja.



KUVIO 1. Yhdyssanojen osuudet hakuaiheissa.

Kuviossa 1 on esitetty sekä SAOL:n että SWETWOL:n tekemät yhdyssanatulkinnat. Luvuista nähdään, että SAOL tunnistaa enemmän yhdyssanoja kuin SWETWOL. Varsinkin, kun otetaan huomioon se, että noin neljäsosa SWETWOL:n tulkitsemista yhdyssanoista on virhetulkintoja. Täytyy kuitenkin myös muistaa se, että SAOL-yhdyssanojen joukossa on kyseenalaisia tapauksia, kun otetaan huomioon tämän tutkimuksen yhdyssanamääritelmä. Kaiken kaikkiaan yhdyssanojen määrän analyysi todistaa sitä tosiasiaa, että yhdyssanat esiintyvät laajamittaisesti kaikissa aineistoissa. Tämän vuoksi onkin tärkeää huomioida yhdyssanojen rooli myös tiedonhaussa.

7.2.2 Yhdyssanatyypit

TAULUKKO 3. Yhdyssanatyypit CLEF2003-hakuaiheissa

	Lukumäärä	Prosenttia	SWETWOL tunnistaa	Prosenttia
Kompositionaaliset	77	60,6	69	89,6
Ei-kompositionaaliset	50	39,4	19	38,0
Yhteensä	127	100,0		
Substantiiviset	107	84,3		
Adjektiiviset	7	5,5		
Yhdyssanaverbit	10	7,9		
Muut tapaukset	3	2,4		
Yhteensä	127	100,0		

Ensimmäiseksi aineistosta on tutkittu kompositionaalisten ja ei-kompositionaalisten yhdyssanojen suhdetta. Yhdyssanatyyppejä on analysoitu SAOL:n tulkinnanmukaisista yhdyssanoista, koska SWETWOL:n tekemissä tulkinnoissa on mukana sanoja, jotka eivät ole yhdyssanoja. Toiseksi aineistosta on analysoitu yhdyssanojen sanaluokat.

Taulukossa 3 on kuvattu CLEF2003-hakuaiheissa esiintyvät yhdyssanatyypit. Kompositionaalisten yhdyssanojen osuus on noin 60 prosenttia, mutta myös ei-kompositionaalisia yhdyssanoja esiintyy jopa 40 prosentin verran. Muistettavaa tässä on SAOL-sanakirjan tulkintojen liberaalisuus. Yhdyssanojen joukossa esiintyy paljon sellaisia yhdyssanoja, joita on vaikea jakaa osiinsa ja jotka näin ollen lasketaan ei-kompositionaalisiksi. Lisäksi taulukossa on esitetty, kuinka monta prosenttia SWETWOL tunnistaa yhdyssanoista. Kuten taulukosta 3 huomataan SWETWOL tunnistaa kompositionaaliset yhdyssanat 90-prosenttisesti. Nämä ovat siis vakiintuneita yhdyssanoja, joiden osat SWETWOL:n on helppo tunnistaa. Ei-kompositionaalisten yhdyssanojen kohdalla SWETWOL tunnistaa vain 38 prosenttia yhdyssanoista. Tässä yhdyssanojen tunnistaminen tarkoittaa sitä, että SWETWOL tekee sanasta analyysin ja sijoittaa yhdyssanan osien väliin joko merkin # tai |.

Taulukossa 3 on lisäksi esitetty eri sanaluokkien osuudet CLEF2003-hakuaiheissa. Jopa 84,3 prosenttia sanoista on substantiivisia. Toiseksi eniten (7,9 prosenttia) esiintyy yhdyssanaverbejä ja kolmanneksi eniten (5,5 prosenttia) adjektiivisia yhdyssanoja. Luokkaan muut kuuluu yhdyssanoja, jotka ovat sanaluokaltaan adverbeja. Näitä esiintyy hakuaiheissa 2,4 prosentin verran.

TAULUKKO 4. Yhdyssanatyypit CLEF2002-hakuaiheissa

	Lukumäärä	Prosenttia	SWETWOL tunnistaa	Prosenttia
Kompositionaaliset	57	55,3	47	82,5
Ei-kompositionaaliset	46	44,7	14	30,4
Yhteensä	103	100,0		
Substantiiviset	77	74,76		
Adjektiiviset	8	7,77		
Yhdyssanaverbit	17	16,50		
Muut tapaukset	1	0,97		
Yhteensä	103	100,00		

Taulukossa 4 esitellään yhdyssanatyypit CLEF2002-hakuaiheissa. Tässä aineistossa kompositionaalisten ja ei-kompositionaalisten yhdyssanojen suhde on tasaisempi 55-45. SWETWOL tunnistaa kompositionaalisista yhdyssanoista 82,5 prosenttia ja ei-kompositionaalisista yhdyssanoista 30 prosenttia. Tässäkin aineistossa substantiivisten yhdyssanojen määrä on suurin (74,76 prosenttia), mutta myös yhdyssanaverbejä on varsin paljon (16,5 prosenttia).

Ahlgren-aineiston yhdyssanatyypit ovat nähtävissä taulukosta 5. Kompositionaalisten ja ei-kompositionaalisten yhdyssanojen suhde on verrattain sama kuin kahdessa muussa aineistossa. SWETWOL tunnistaa kompositionaaliset yhdyssanat 84-prosenttisesti. Ei-kompositionaalisistakin yhdyssanoista SWETWOL:lle on tuttuja jopa 53 prosenttia. Myös tässä aineistossa substantiivisia yhdyssanoja esiintyy eniten (81 prosenttia), yhdyssanaverbejä toiseksi eniten noin 12 prosenttia ja adjektiivisia yhdyssanoja noin 5 prosenttia.

TAULUKKO 5. Yhdyssanatyypit Ahlgren-hakuaiheissa

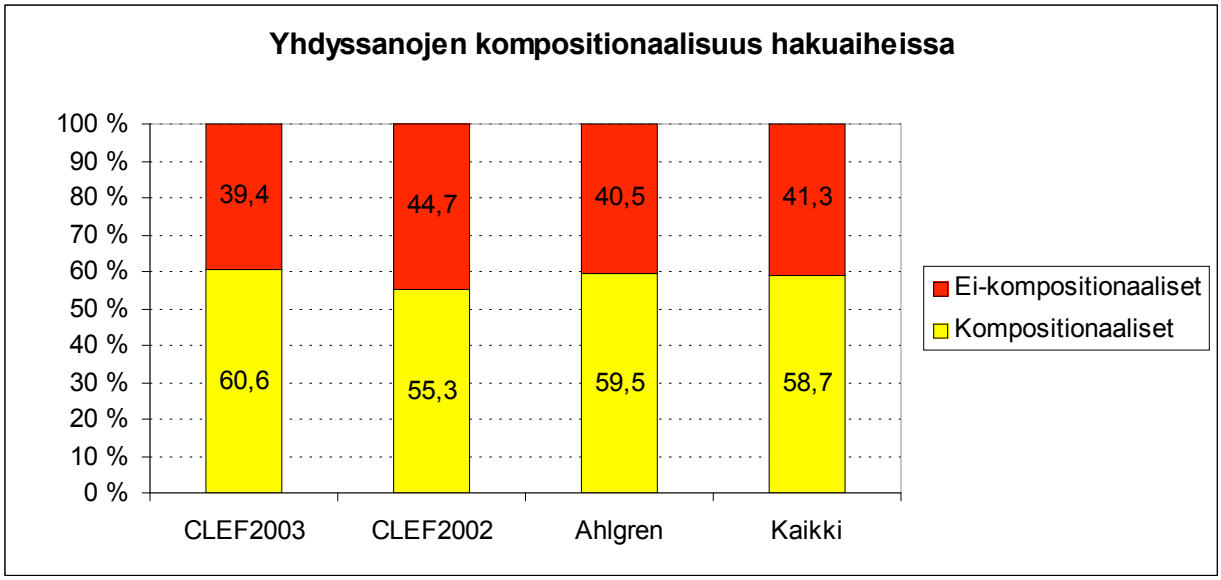
	Lukumäärä	Prosenttia	SWETWOL tunnistaa	Prosenttia
Kompositionaaliset	69	59,5	58	84,1
Ei-kompositionaaliset	47	40,5	25	53,2
Yhteensä	116	100,0		
Substantiiviset	94	81,03		
Adjektiiviset	6	5,17		
Yhdyssanaverbit	14	12,07		
Muut tapaukset	2	1,72		
Yhteensä	116	100,0		

TAULUKKO 6. Yhdyssanatyypit kaikissa hakuaiheissa yhteensä

	Lukumäärä	Prosenttia	SWETWOL tunnistaa	Prosenttia
Kompositionaaliset	203	58,7	174	85,7
Ei-kompositionaaliset	143	41,3	58	40,6
Yhteensä	346	100,0		
Substantiiviset	278	80,35		
Adjektiiviset	21	6,07		
Yhdyssanaverbit	41	11,85		
Muut tapaukset	6	1,73		
Yhteensä	346	100,0		

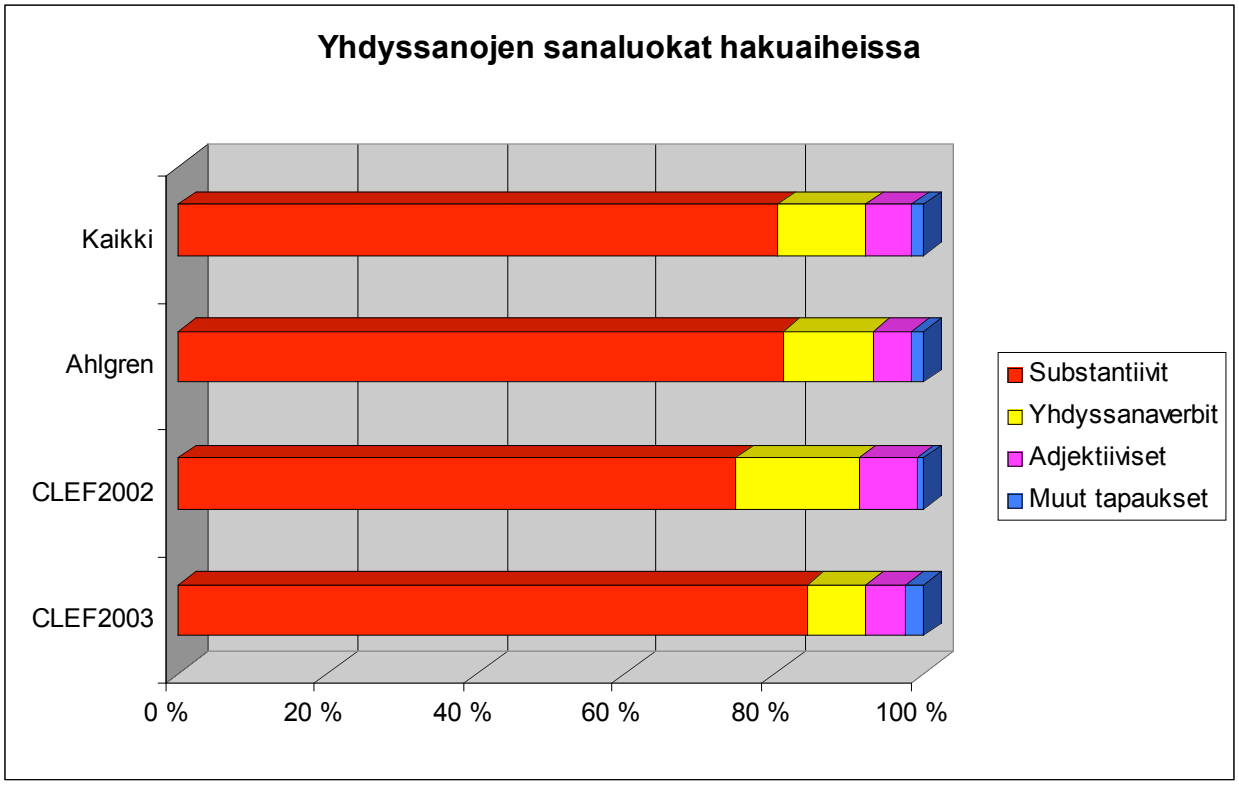
Taulukko 6 esittää yhdyssanatyyppien jakaumat kaikissa hakuaiheissa yhteensä. Kompositionaalisia yhdyssanoja on noin 59 prosenttia kaikista yhdyssanoista. SWETWOL tunnistaa näistä 86 prosenttia. Ei-kompositionaalisten yhdyssanojen osuus on noin 41 prosenttia, ja niistä SWETWOL tunnistaa 40 prosenttia. Kompositionaalisuuden vertailu kaikissa hakuaiheissa on myös esitetty kuviossa 2. Karkeasti ottaen kompositionaalisten ja ei-kompositionaalisten yhdyssanojen välinen suhde näyttäisi olevan aineistossa 60-40. Ei-kompositionaalisten yhdyssanojen joukossa on mukana monia luvussa 7.1 esiteltyjä ongelmatapauksia. Nämä ovat myös sellaisia yhdyssanoja, joita SWETWOL ei tunnista. Koska kompositionaalisia yhdyssanoja on enemmän kuin ei-kompositionaalisia yhdyssanoja, voitaisiin ajatella, että yhdyssanojen morfologinen käsittely on hyödyllistä ruotsin kielessä. Ei-kompositionaalisten yhdyssanojen osuuskin on varsin suuri, joten valikoiva osittaminen saattaa olla ratkaisu ongelmiin. Tosin on muistettava, että ei-kompositionaalisten joukossa on paljon sanoja, joita SWETWOL ei tunnista ja jotka eivät täysin

täytä yhdyssanamääritelmää. Yhdyssanojen osittaminen ei myöskään ole tarkoituksenmukaista muiden kuin substantiivisten yhdyssanojen kohdalla, koska osituksen tuloksena syntyvät yhdysosat eivät ole merkityksellisiä hakuavaimia. Esimerkiksi yhdyssanaverbit muodostuvat usein partikkelista ja verbistä (**under-visa**).



KUVIO 2. Yhdyssanojen kompositionaalisuus hakuaiheissa.

On kuitenkin otettava huomioon, että analyysi ei kerro koko totuutta. Tutkimuksen taustaoletuksena on, että kompositionaalisten yhdyssanojen osittaminen on hyödyllisempää kuin ei-kompositionaalisten yhdyssanojen. Tiedonhaun näkökulmasta kaikkia kompositionaalisiakaan yhdyssanoja ei kuitenkaan kannata jakaa osiin. Jos yhdyssanojen osat ovat esimerkiksi todella yleisiä, ne eivät ole hyviä hakuavaimia, esimerkiksi **rik-dags-hus**. Sanan yleisyys dokumentissa vaikuttaa sen hyödyllisyyteen hakuavaimena, ja tähän seikkaan kompositionaalisten ja ei-kompositionaalisten yhdyssanojen tunnistaminenkaan ei voi vaikuttaa.



KUVIO 3. Yhdyssanojen sanaluokat hakuaiheissa.

Kuten kuviosta 3 huomataan kaikissa hakuaiheissa on eniten substantiivisia yhdyssanoja ja toiseksi eniten yhdyssanaverbejä. CLEF2002 poikkeaa muista aineistoista suuremman yhdyssanaverbien määränsä suhteen. Jos ajatellaan sanaluokkainformaatiota tiedonhaun näkökulmasta, aineisto sisältää eniten substantiivisia yhdyssanoja ja substantiivit ovat myös hakuavaimina enemmistössä. Yhdyssanojen osittamisen kannalta yhdyssanaverbien esiintyminen ei ole suuri ongelma, koska a) niitä esiintyy aineistossa suhteellisen vähän ja b) yhdyssanojen käsittelyyn käytettävä ohjelma SWETWOL tunnistaa niistä vain pienen osan. Tämän vuoksi kyselyihin ei välttämättä tule haun kannalta hyödyttömiä yhdyssanaverbien osia.

7.3 Yhdyssanojen määrä ja tyypit dokumenteissa

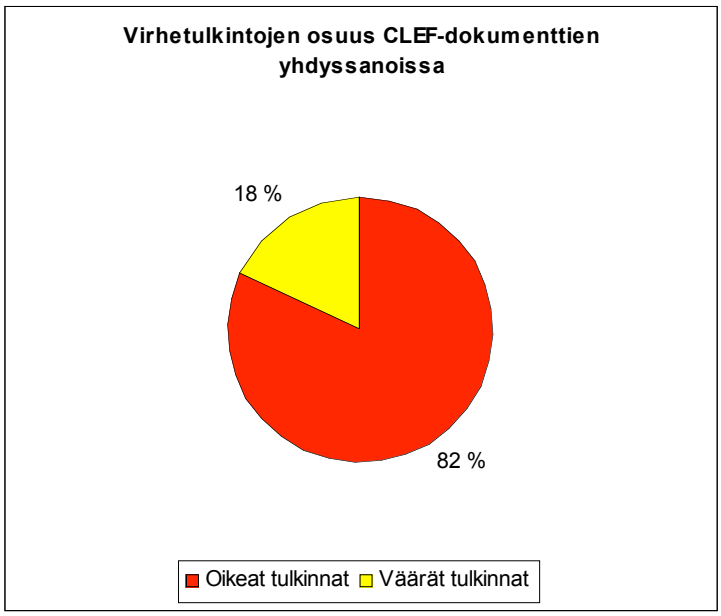
Yhdyssanojen määrää dokumenteissa tutkitaan aineiston koon vuoksi otoksen avulla. Selvyyden vuoksi aineisto on jaettu kahtia CLEF-aineistoon (sisältää vuoden 2002 ja 2003 aineistot) ja Per Ahlgrenin kokoelmaan. Aineistoista on poimittu yhteensä 200 dokumenttia. Yhdyssanatyypeistä tutkitaan dokumenttien osalta pelkästään kompositionaalisten ja ei-kompositionaalisten yhdyssanojen välistä suhdetta.

7.3.1 CLEF-aineiston dokumentit

TAULUKKO 7. CLEF-dokumenttien sanat

	Lukumäärä	Prosenttia
Ei-yhdyssanat	18026	86,17
Yhdyssanat	2854	13,64
Vajaat rivit	39	0,19
Yhteensä	20919	100,0

Taulukossa 7 on esitetty yhdyssanojen määrä CLEF-aineiston dokumenteissa. Yhdyssanojen osuus on 13,6 prosenttia, mikä on hakuaiheiden tulosten mukaisesti hieman enemmän kuin Hedlundin saama tulos 9,8 prosenttia 100 000 sanan aineistoa koskevassa arviossaan. Dokumenttien yhdyssanamäärä on lähempänä hakuaiheanalyysin SWETWOL-tulkintojen määrää. SAOL-yhdyssanoja oli hakuaiheanalyysissä CLEF2003-aineiston osalta jopa 17 prosenttia. Täytyy kuitenkin myös muistaa, että CLEF2002-hakuaiheiden osalta yhdyssanamäärä oli sekä SAOL:n (14,5 prosenttia) että SWETWOL:n osalta (noin 11 prosenttia) muita aineistoja pienempi. Myös SWETWOL:n käyttäminen apuna yhdyssanojen laskemisessa vaikuttaa määrään. SWETWOL:n tekemiin tulkintoihin sisältyy virhetulkintoja ja toisaalta SWETWOL ei välttämättä tunnista kaikkia yhdyssanoja. Vajaat rivit taulukossa 7 tarkoittaa aineiston sisältämiä vajaita yhdyssanatulkintoja. Esimerkiksi aineistossa on yksittäisiä sanoja, joiden edessä on yhdysmerkki (-**direktivet**) ja joista yhdyssanan etuosa on kadonnut. Näitä sanoja ei ole laskettu mukaan yhdyssanoiksi.



KUVIO 4. Virhetulkinnat CLEF-dokumenttien yhdyssanoissa.

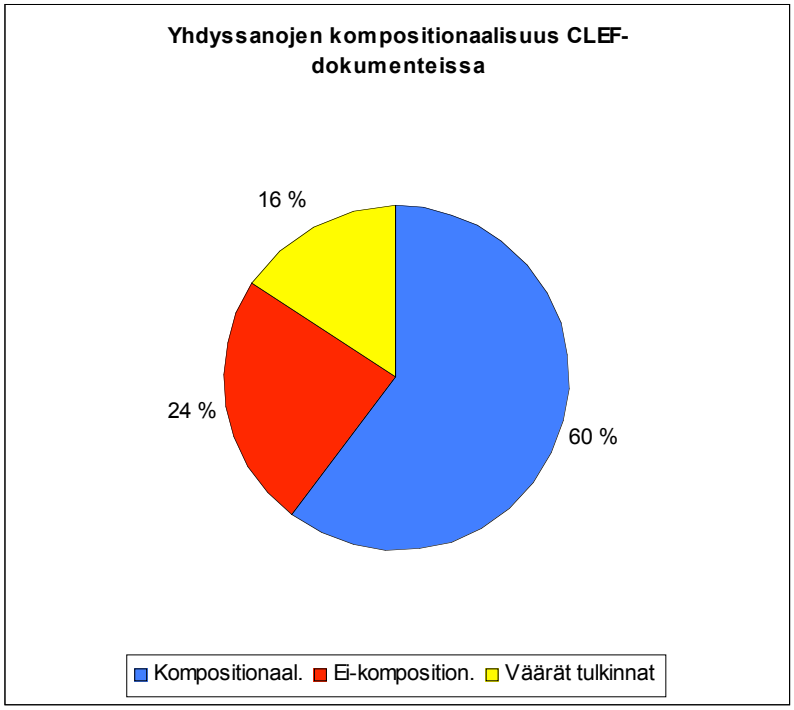
Kuviosta 4 käy ilmi SWETWOL:n tekemät tulkinnat. Yhdyssanatulkinnoista 82 prosenttia on oikeita ja loput 18 prosenttia vääriä tulkintoja. Vääriä tulkintoja voi olla kahdenlaisia. SWETWOL on voinut tulkita yhdyssanaksi sanan, joka ei ole sitä. Esimerkiksi sanoissa **budgeten** tai **ligger**. Toisaalta väärien tulkintojen joukossa voi olla sanoja, jotka ovat yhdyssanoja, mutta jotka SWETWOL on tulkinnut väärällä tavalla. Esimerkiksi sanassa **uppåt** tulkinta on **upp-äta**. Muita esimerkkejä SWETWOL:n tekemistä virhetulkinnoista ovat seuraavat: **blygsam** (blyg-simma), **riktiga** (rik-tiga) ja **underskott** (under-sko).

Yhdyssanatulkintojen joukossa on myös lukuisia erisnimiä, joista osa voidaan tulkita yhdyssanoiksi (esimerkiksi **Skeppsholmen**, **Ryssland**), osa virhetulkinnoiksi (esimerkiksi **Ingrid**, **Mattias**). Gunnarsonin ja Peterssonin (2005) tutkimuksessa on todettu, että yhdyssanojen osittaminen ei ole hyödyllistä erisnimien kohdalla. Yhdyssana-aineisto sisältää kuitenkin lukuisia erisnimiyhdyssanoja, jotka myös SWETWOL on tulkinnut yhdyssanoiksi. Erisnimien lisäksi yhdyssanatulkintojen joukossa on tilapäisiä yhdyssanoja (**wallenbergmålet**) ja leksikaalistuneita yhdyssanoja (**måndagen**). Joukossa on myös ongelmallisia ja vaikeasti tulkittavia sanoja, (esimerkiksi full-, lös-, fri- ja bar-loppuiset sanat). SAOL- ja SAOB-sanakirjoja apuna käyttäen tulkitsin nämä sanat yhdyssanoiksi.

TAULUKKO 8. Kompositionaaliset yhdyssanat (CLEF)

	Lukumäärä	Prosenttia
Kompositionaal.	1723	60,4
Ei-komposition.	676	23,7
Väärät tulkinnat	455	15,9
Yhteensä	2854	100

Taulukossa 8 ja kuviossa 5 on esitetty kompositionaalisten ja ei-kompositionaalisten yhdyssanojen jakaumat CLEF-dokumenttien osalta. Kompositionaalisten yhdyssanojen osuus on kuten hakuaiheissakin 60 prosentin luokkaa. Loput 40 prosenttia jakautuu ei-kompositionaalisten yhdyssanojen ja väärien tulkintojen kesken siten, että ei-kompositionaalisia yhdyssanoja on noin 24 prosenttia. Väärien tulkintojen joukossa ovat sanat, jotka SWETWOL on tulkinnut yhdyssanoiksi, mutta jotka eivät ole yhdyssanoja ja joita ei näin ollen voi tyypitellä kumpaankaan yhdyssanatyyppiin. Virhetulkintojen määrä on erilainen (kuvio 4 ja taulukko 8), koska SWETWOL:n tekemistä virhetulkinnoista osa on voinut olla yhdyssanoja, jotka SWETWOL on vain tulkinnut väärin. Aiemmin esitetty väärä tulkinta **uppåt** on esimerkiksi nyt laskettu mukaan kompositionaalisiin yhdyssanoihin, ja se ei ole esillä taulukon 8 virhetulkinnoissa.



KUVIO 5. Yhdyssanojen kompositionaalisuus CLEF-dokumenteissa.

Esimerkkejä aineiston kompositionaalisista yhdyssanoista ovat **valbudget**, **arbetsmarknad**, **barnfamilj** ja **diamantsmuggling**. Ei-kompositionaalisten yhdyssanojen joukossa on substantiiveja (**halvlek, riksdag, bostad, område, morot, medlem, översyn, måndag, middag, vardag, trettonhelgen**), verbejä (**påstå, uppge**), adjektiiveja (**konstgjord, sannolik**) sekä erisnimiä (**Nederländerna, Småland**). Aineiston yhdyssanojen joukosta pystyy myös erottamaan sanoja, jotka aiheuttavat ongelmia homografian vuoksi. Esimerkiksi yhdysosana ongelmallinen sana on **rik**, koska se voidaan ymmärtää sekä sanaksi **rik** 'rikas' että sanaksi **rike** 'valtakunta'. Esimerkkejä tästä yhdysosasta muodostetuista yhdyssanoista ovat muun muassa **riksförbund** ja **riksåklagare**. Lisäksi aineistossa esiintyy virhetulkintoja, joiden yhdysosana **rik** toimii: **riktiga** ja **fredrik**. Toinen esimerkki homografiasta on sana **under** (merkitykseltään 'ihme' tai 'alla'), joka esiintyy erityisesti yhdyssanaverbien yhdysosana (**undergå**, **undervisa**, **underskatta**) ja yhdysverbeistä muodostetuissa substantiiveissa (**underlag**). Myös **för**-sanan homografisuus ja runsas esiintyminen näkyy aineistossa.

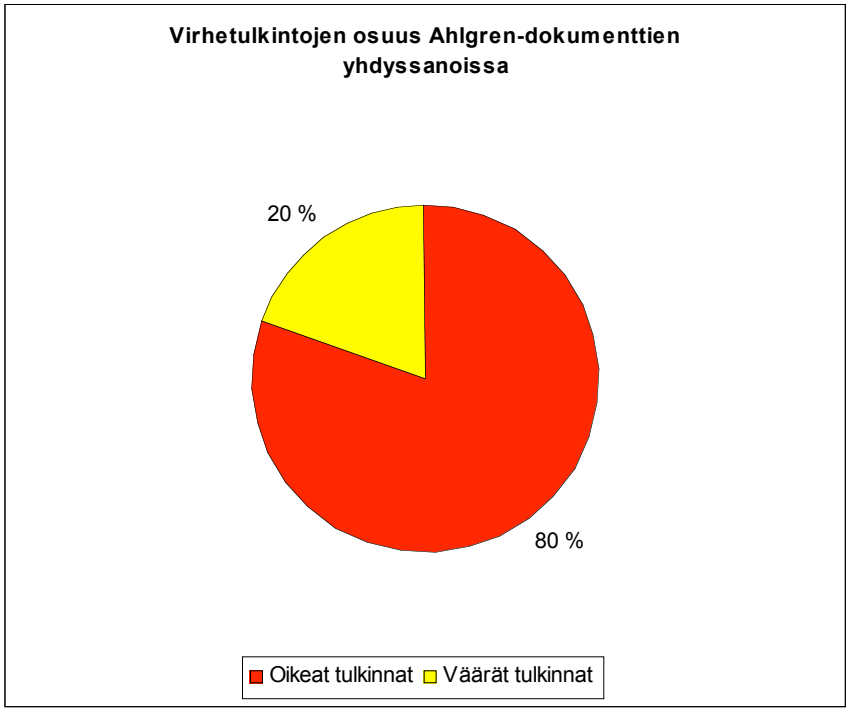
Näiden lisäksi aineistossa on erotettavissa SWETWOL:n virhetulkintoja aiheuttavia sanaaineksia. Tällaisia ovat esimerkiksi ainekset **ko**, **kor**, **sko**, **skor**, joiden vuoksi aiheutuvat virhetulkinnat **styrkor**, **korgen**, **väskor**, **japanskorna** ja **svenskorna**. Aines **port** aiheuttaa myös virheellisiä tulkintoja (esimerkiksi **rapport**, **transport**). Ainekset **lik**, **lig** ja **liga** aiheuttavat vir-

hetulkintoja (esimerkiksi **publik**). Samoin ainekset **bet** ja **pet** (virhetulkinta **jobbet**, **loppet**), **ge** (**roger**), **sten** (**vinsten**) sekä **bar** (**klubbarna**).

Aineisto jakautuu kahtia yhdyssanatulkintoihin ja muihin sanoihin, joita SWETWOL ei ole tulkinnut yhdyssanaksi. Näiden muiden sanojen joukossa näyttäisi kuitenkin olevan yhdyssanoja. Ensimmäisen tuhannen sanan joukossa on 21 yhdyssanaa. Ne jakautuvat tilapäisiin yhdyssanoihin (**enprocentmålet, tt-profil**), yhdyssanaverbeihin (**uppnå**, **anslå**), verbeistä muodostettuihin yhdyssanoihin (**nedskärningar**, **sysselsättning**), erisnimiin (**friggebo, Svensson**) sekä leksikaalistuneisiin yhdyssanoihin (**torsdagen, allmän**). Seuraavan tuhannen sanan joukossa yhdyssanoja on 39 ja joukossa on tilapäisiä yhdyssanoja (**spetsnaz-soldater**, **fyrhjulingar**), leksikaalistuneita yhdyssanoja (**torsdagen**, **densamma**, **enstaka**, **enskild**, **gårdagen**), yhdyssanaverbejä (**utväxla**, **utgå**, **föreslå**) sekä monia verbeistä johdettuja substantiiveja (**nedgång**, **uppgång**). Seuraavan tuhannen sanan joukossa on myös joukko yhdyssanoja (jopa 40 kappaletta), joiden joukossa on tilapäisiä yhdyssanoja (**andraåk**, **vip-läktaren**, **os-arenorna**), verbeistä muodostettuja yhdyssanoja (**uppgång**, **antal**, **förbjuden**, **förbud**, **påklädd**, **övervältigad**) sekä vakiintuneita ja leksikaalistuneita yhdyssanoja (**allmän**, **spörsmål**, **detsamma**, **onsdagen**, **förmiddagen**). Viikonpäivien osalta mielenkiintoista on se, että SWETWOL tulkitsee yhdyssanaksi muun muassa sanan **måndagen**, mutta **torsdagen** on ei-yhdyssanojen listalla.

7.3.2 Ahlgren-dokumentit

Taulukosta 9 käyvät ilmi ei-yhdyssanojen ja yhdyssanojen jakaumat Ahlgren-aineiston dokumenteissa. Yhdyssanoja aineistossa on 12,45 prosenttia. Yhdyssanojen määrä CLEF-dokumenteissa on 13,6 prosenttia, joten määrä on suhteellisen sama eli hieman enemmän kuin Hedlundin arvio 9,8 prosentista. Ahlgren-hakuaiheissa yhdyssanojen määrä oli SAOL:n osalta 17,4 prosenttia ja SWETWOL-tulkintojen osalta 15,6 prosenttia. Vajaita rivejä (**-årig**) esiintyy tässäkin aineistossa, ja ne on mainittu erikseen taulukossa 9. CLEF-dokumenttien tavoin SWETWOL:n tekemien virhetulkintojen osuus on noin 20 prosenttia, mikä käy ilmi kuviosta 6.



KUVIO 6. Virhetulkinnat Ahlgren-dokumenttien yhdyssanoissa.

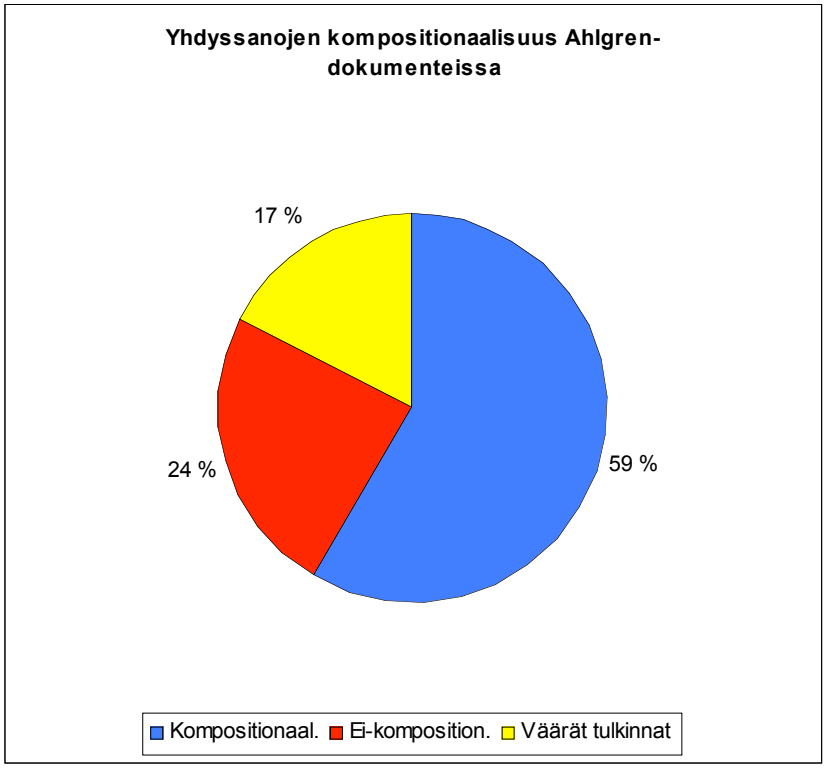
TAULUKKO 9. Ahlgren-dokumenttien sanat

	Lukumäärä	Prosenttia
Ei-yhdyssanat	26361	87,32
Yhdyssanat	3760	12,45
Vajaat rivit	69	0,23
Yhteensä	30190	100,00

Kuten taulukosta 10 ja kuviosta 7 käy ilmi Ahlgrenin aineiston yhdyssanoista lähes 60 prosenttia on kompositionaalisia, mikä vahvistaa sekä hakuaiheiden että CLEF-dokumenttien osalta saatuja tuloksia. Loppuosa jakautuu 24 prosentin ei-kompositionaalisten yhdyssanojen osuuteen ja 17 prosentin väärien tulkintojen osuuteen. Samoin kuin CLEF-dokumenteissa joukossa on paljon erisnimiyhdyssanoja. Mukana on niin vääriä tulkintoja (**Hyllinge**, **James**) kuin yhdyssanamuodossa olevia erisnimiäkin (**Eksjö**, **Skottland**, **Eliasson**).

TAULUKKO 10. Kompositionaaliset yhdyssanat (Ahlgren)

	Lukumäärä	Prosenttia
Kompositionaal.	2197	58,4
Ei-komposition.	908	24,1
Väärät tulkinnat	655	17,4
Yhteensä	3760	100,0



KUVIO 7. Yhdyssanojen kompositionaalisuus Ahlgren-dokumenteissa.

Esimerkkejä aineiston kompositionaalisista yhdyssanoista ovat yhdyssanaverbi **innehålla**, substantiiviyhdyssana **hamnstad** ja erisnimiyhdyssana **Nordamerika**. Esimerkkejä ei-kompositionaalisista yhdyssanoista ovat **farvatten**, **styvfar**, **gudfader**, **samvete**, **kyrkoherden**, **grönsak**, **vitsord**, **pannkaka**, **bakläxa** ja **frimärke**. Aineiston yhdyssanojen joukosta pystyy myös erottamaan sanoja, jotka aiheuttavat ongelmia homografian vuoksi. Esimerkiksi yhdysosana ongelmallinen sana on **rik**, koska se voidaan ymmärtää sekä sanaksi **rik** 'rikas' että sanaksi **rike** 'valtakunta'. Esimerkkejä tästä yhdysosasta muodostetuista yhdyssanoista ovat muun muassa **riksfinalen** ja **rikskriminal**. Esimerkkejä virhetulkinnoista, jotka sisältävät osan **rik**, ovat **Fredrik**, **Henrik**, **tallrik** ja **riktig**. Toinen esimerkki homografiasta on sana **under**, joka esiintyy aineistossa yhdyssanaverbeissä ja niistä muodostetuissa substantiiveissa (**underlätta**, **undersökning**, **underhålla**). Myös homografinen sana **för** esiintyy aineiston yhdyssanoissa (**framför**, **jämföra**, **genomföra**). Ahlgren-dokumenteissa esiintyy myös samankaltaisia virhetulkintoja aiheuttavia aineksia kuin jo mainittiin CLEF-dokumenttien yhteydessä (esimerkiksi **lik**, **liga**, **ko** ja **sko**). Aineistossa on myös paljon epäselviä yhdyssanatapauksia. Esimerkiksi full-päätteiset yhdyssanat (**gåtfull**, **smakfull**) ja yhdyssana-adjektiivi **vädermässig** sekä yhdyssanan sisältävä johdos **medlemskap**. Nämä tapaukset ovat SAOL- ja SAOB-sanakirjojen mukaan yhdyssanoja.

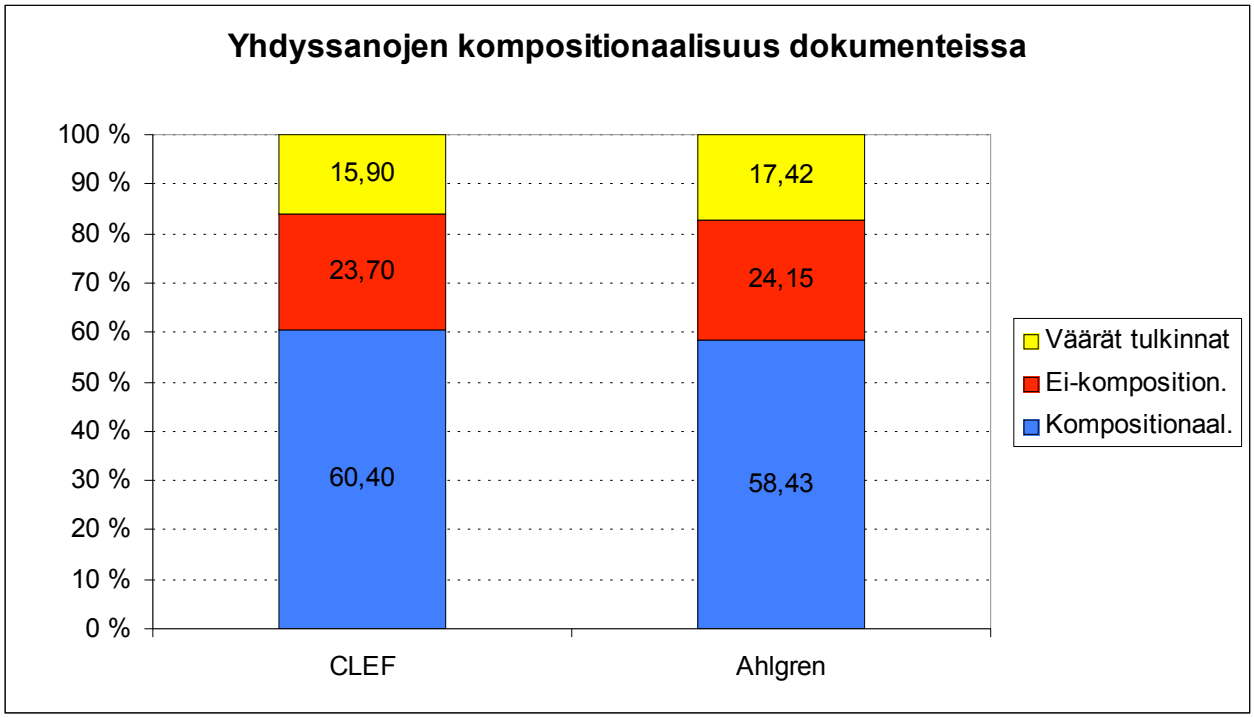
Aineisto jakautuu kahtia yhdyssanatulkintoihin ja muihin sanoihin, joita SWETWOL ei ole tulkinnut yhdyssanaksi. Näiden muiden sanojen joukossa on kuitenkin yhdyssanoja. Ensimmäisen tuhannen sanan joukossa on 34 yhdyssanaa. Joukossa on erisnimiä (**Johansson**, **Olsson**, **Karlsson**), yhdyssanaverbejä ja niistä muodostettuja sanoja (**upptäcka**, **uppdaga**, **utskriven**, **uteblíven**), viikonpäivien nimiä (**onsdag**) sekä substantiiveja, joita SWETWOL ei tunnista (**nattetid**, **utsida**). Seuraavan tuhannen sanan joukossa esiintyy 29 yhdyssanaa. Ne jakautuvat erisnimiin (**Einarsson**, **Ängelholm**), yhdyssanaverbeihin ja niistä muodostettuihin sanoihin (**utveckla**, **inledande**, **förslag**, **indrag**) ja tilapäisiin yhdyssanoihin (**os-målvakt**, **vik-kassen**, **os-hjälte**, **assidomänaktie**). Seuraavan tuhannen sanan joukossa on 32 yhdyssanaa. Esimerkiksi yhdyssanoina esiintyy erisnimiä (**Fredriksdal**, **Djursholm**, **Övertorneå**), yhdyssanaverbejä ja niistä muodostettuja sanoja (**inleda**, **utforma**, **föreslå**, **utvärdering**, **redovisning**), viikonpäivien nimiä (**fredag**, **torsdag**), adverbejä (**framemot**, **huruvida**) sekä tilapäisiä yhdyssanoja (**jonstorpsskolan**, **magneklint**, **vikenskolan**, **korp-badminton**).

7.3.3 CLEF vs. Ahlgren

TAULUKKO 11. Yhdyssanojen määrä (%) dokumenteissa.

	CLEF	Ahlgren
Ei-yhdyssanat	86,17	87,32
Yhdyssanat	13,64	12,45
Vajaat rivit	0,19	0,23
Yhteensä	100	100

Taulukossa 11 on esitetty vertailu yhdyssanojen määrästä aineistojen dokumenteissa. Yhdyssanojen määrä on molemmissa aineistoissa samankaltainen. Hakuaiheiden ja dokumenttien perusteella saadut tulokset viittaavat siihen, että yhdyssanojen määrä on ruotsin kielessä hieman Hedlundin esittämää lukua suurempi. Tiedonhaun näkökulmasta on merkittävää, että näinkin suuri osa tekstiaineiston sanoista on yhdyssanoja. Seuraavassa luvussa pyritään selvittämään keinoja yhdyssanojen käsittelemiseen tiedonhaussa.



KUVIO 8. Vertailu: yhdyssanojen kompositionaalisuus dokumenteissa.

Kuviossa 8 on vertailtu yhdyssanojen kompositionaalisuuden osuuksia CLEF- ja Ahlgren-dokumenteissa. Molemmissa aineistoissa kompositionaalisuus on noin 60 prosenttia, aivan kuten hakuaiheiden analyysissä jo on todettu. Ei-kompositionaalisten osuus on samoin molemmissa aineistoissa 24 prosentin luokkaa. Hakuaiheissa ei-kompositionaalisten yhdyssanojen osuus oli suurempi, noin 40 prosenttia. Hakuaiheiden analyysissä ei ollut mukana virhetulkintoja, koska analyysin kohteena olivat SAOL-sanakirjan tekemät yhdyssanatulkinnat. Kompositionaalisten yhdyssanojen enemmistöosuus antaisi viitteitä siitä, että yhdyssanojen morfologinen käsittely voisi ruotsin kielen osalta olla tarkoituksenmukaista. Seuraavassa luvussa onkin tarkoituksena selvittää, miten yhdyssanojen osittaminen ja yhdyssanojen eliminoiminen vaikuttavat hakutuloksiin ruotsin kielessä.

8 YHDYSSANOJEN MORFOLOGISEN KÄSITTELYN VAIKUTUS HAKUTULOKSIIN

Tässä luvussa esittelen laboratoriotutkimuksen tulokset. Tutkimuksessa on vertailtu eri kyselynmuodostustapoja ruotsin kielessä, ja vastausta haetaan siihen, kannattaako yhdyssanoja käsitellä

ruotsinkielisessä tiedonhaussa. Tutkimuksessa on käytössä kolme eri aineistoa, CLEF2003, CLEF2002 ja Per Ahlgrenin kokoelma. Tiedonhakuja tehdään sekä yhdyssanojen osalta ositettuun ja eliminoituun hakemistoon (käytetään termiä eliminoitu kanta) että yhdyssanoiltaan ositettuun, mutta eliminoimattomaan hakemistoon (käytetään termiä eliminoimaton kanta). Kutakin hakemistoa kohden haetaan kolmella erilaisella kyselysarjalla:

TAULUKKO 12. Tutkimuksen kyselysarjat

ELIMINOITU KANTA		
Perusmuotoinen, yhdyssanoja ei ositettu = **Ei ositettu**	vs.	Perusmuotoinen, yhdyssanat ositettu, ei eliminoitu = **Eliminoimaton**
	vs.	Perusmuotoinen, yhdyssanat ositettu, eliminoitu = **Eliminoitu**
ELIMINOIMATON KANTA		
Perusmuotoinen, yhdyssanoja ei ositettu = **Ei ositettu**	vs.	Perusmuotoinen, yhdyssanat ositettu, ei eliminoitu = **Eliminoimaton**
	vs.	Perusmuotoinen, yhdyssanat ositettu, eliminoitu = **Eliminoitu**

Kustakin kyselysarjasta käytetään sekä rakenteista että rakenteetonta versiota.

8.1 Rakenteiset kyselysarjat

Ensimmäiseksi esittelen rakenteisilla eli #syn- ja #uw20-operaattoreita hyödyntävillä kyselysarjoilla saadut tulokset kolmen eri aineiston osalta.

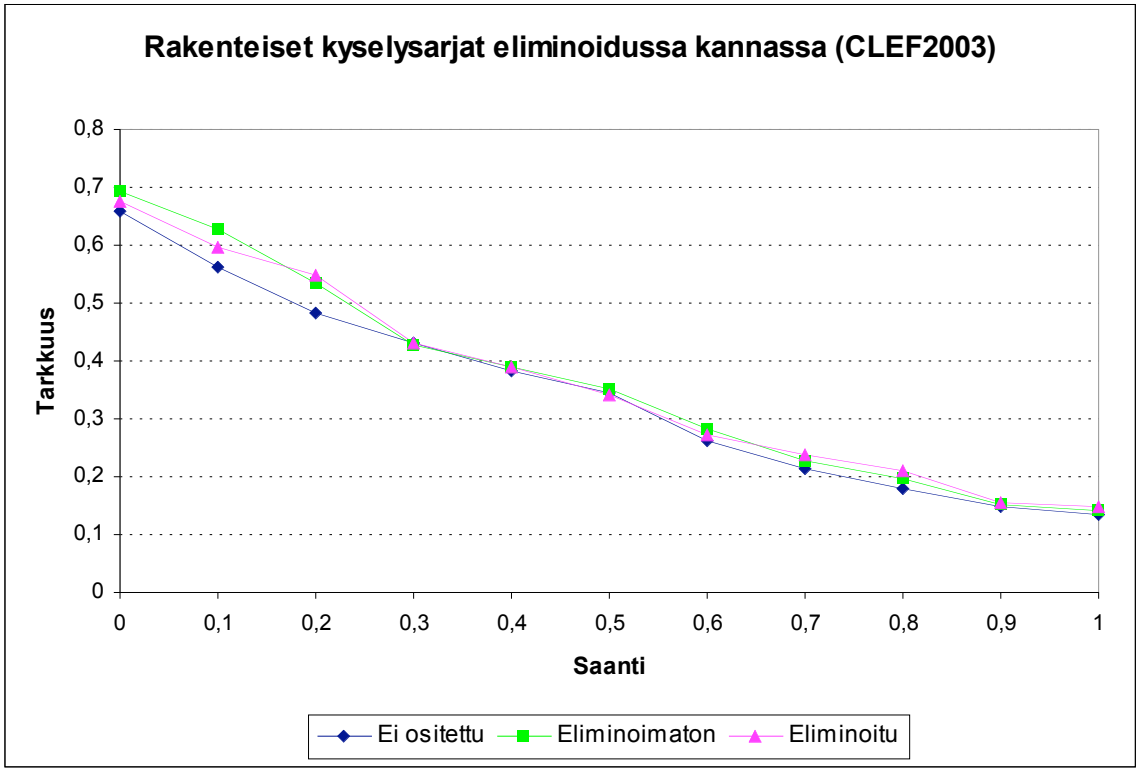
8.1.1 CLEF2003-kyselysarjoilla saadut tulokset

Taulukossa 13 on esitetty vertailtujen rakenteisten kyselysarjojen keskimääräiset ei-interpoloidut tarkkuusarvot prosentteina CLEF2003-aineiston osalta. Ensiksikin taulukossa esitetään prosenttiyksikköinä ei ositetun ja yhdyssanoiltaan ositetun, mutta eliminoimattoman kyselysarjan ero. Eliminoidun kannan osalta perusmuotoisen, yhdyssanoiltaan osittamattoman kyselysarjan tarkkuus on 33,16 prosenttia, kun taas yhdyssanoiltaan ositetun kyselysarjan tarkkuus on 35,01. Yhdyssanojen osittaminen parantaa tarkkuusarvoa siis 1,85 prosenttiyksikköä. Toiseksi taulukossa

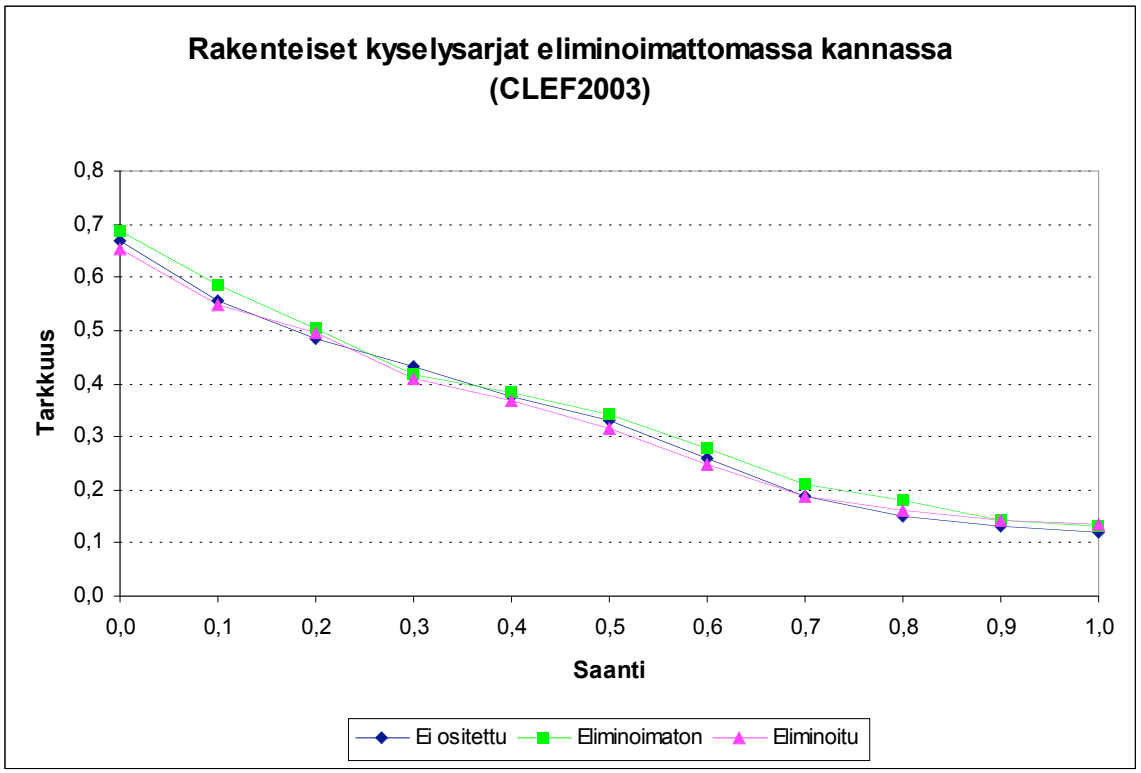
esitetään ei ositetun kyselysarjan sekä ositetun ja eliminoidun kyselysarjan välinen ero prosenttiyksikköinä. Myös yhdyssanojen eliminoiminen parantaa tarkkuutta eliminoidussa kannassa 1,82 prosenttiyksikön verran. Eliminoimattomassa kannassa yhdyssanojen osittaminen ilman eliminoimista parantaa tarkkuutta 1,7 prosenttiyksikköä. Eliminoimattoman kannan osalta perusmuotoinen, yhdyssanoiltaan osittamaton kyselysarja saavuttaa 31,94 prosentin tarkkuuden, kun taas yhdyssanoiltaan ositettu ja eliminoitu kyselysarja saa 31,70 prosentin tarkkuuden. Menetelmien välillä ei ole suurta eroa suuntaan tai toiseen, vaan yhdyssanojen eliminointi huonontaa tarkkuusarvoa ainoastaan 0,24 prosenttiyksikköä. Vertailtujen menetelmien välillä ei havaittu tilastollisissa testeissä merkitseviä eroja.

TAULUKKO 13. Keskimääräiset ei-interpoloidut tarkkuusarvot (%) rakenteisten CLEF2003-kyselysarjojen osalta

	ELIMINOITU KANTA	ELIMINOIMATON KANTA
Ei ositettu	33,16	31,94
Ositettu, eliminoimaton	35,01	33,64
Ei ositetun ja ositetun ero (%-yks.)	1,85	1,7
Ositettu, eliminoitu	34,98	31,7
Ei ositetun ja eliminoidun ero (%-yks.)	1,82	-0,24



KUVIO 9. Rakenteisilla CLEF2003-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa.



KUVIO 10. Rakenteisilla CLEF2003-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa.

Kuviossa 9 on saanti-tarkkuus -käyrä CLEF2003-aineiston rakenteisilla kyselysarjoilla saaduista tarkkuusarvoista eliminoidussa kannassa. Kuten saanti-tarkkuus -käyrästä käy ilmi erot eri menetelmien välillä ovat varsin pienet. Käyrän alkupäässä eliminoimaton ja eliminoitu kyselysarja menestyvät ei ositettua kyselysarjaa paremmin. Kuvio 10 kuvaa CLEF2003-aineiston rakenteisia kyselysarjoja eliminoimattomassa kannassa, jossa menetelmien väliset erot ovat pieniä.

TAULUKKO 14. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn (CLEF2003, rakenteiset kyselyt)

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	19	20
Huonontunut	21	15
Sama	8	13
Ei yhdyssanoja	6	6
Yhteensä	54	54
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	20	16
Huonontunut	20	17
Sama	8	15
Ei yhdyssanoja	6	6
Yhteensä	54	54

Saanti-tarkkuus -käyrät ja tarkkuuksien keskiarvo esittävät keskimääräiset tulokset kaikkien kyselyiden osalta. CLEF2003-kyselysarjojen osalta tutkittiinkin myös hakuaihekohtaisia tarkkuuksia. Taulukossa 14 on esitetty tarkkuuksien muutos yhdyssanojen osittamisen ja yhdyssanojen eliminoimisen myötä. Hakuaiheita on yhteensä 54 kappaletta, ja kuudessa hakuaiheessa ei ole lainkaan yhdyssanoja. Eliminoidussa kannassa yhdyssanojen osittaminen parantaa keskimääräistä tarkkuusarvoa 19 hakuaiheessa ja huonontaa 21 hakuaiheessa, kun taas kahdeksassa hakuaiheessa tarkkuus säilyy samana. Yhdyssanojen eliminoiminen puolestaan säilyttää osittamattomaan kyselyyn verrattuna tarkkuuden samana jopa 13 hakuaiheessa, parantaa tarkkuutta 20 hakuaiheessa ja huonontaa tarkkuutta 15 hakuaiheessa. Eliminoimattomissa kyselyissä tarkkuutta huonontaa yhdysosien runsas määrä verrattuna yhdyssanoiltaan eliminoituihin kyselyihin. Eliminoimattomassa kannassa tarkkuuksien muutokset ovat samansuuntaisia. Erityisesti eliminoidun kyselysarjan osalta tarkkuuden muutokset jakautuvat tasaisesti. Tarkkuuden huonontumista tapahtuu 17 hakuaiheessa, parantumista 17 hakuaiheessa ja tarkkuus pysyy samana 15 hakuaiheessa.

Yksittäisiä hakuaiheita ja niistä muodostettuja kyselyitä tutkimalla on mahdollista löytää selityksiä siihen, missä tilanteissa yhdyssanojen käsittely on hyödyllistä. Esimerkiksi hakuaiheen 142

osalta molemmat yhdyssanojen käsittelymenetelmät huonontavat tarkkuutta. Hakuaihe sisältää yhdyssanan **riksdagshuset**, jonka osittaminen ei ole hyödyllistä. Eliminoimattomassa kyselyssä on mukana niinkin yleiset yhdysosat kuin **rik**, **dag** ja **hus**. Myös eliminoitu kysely on tarkkuudeltaan huono, vaikka siinä ei esiinnykään aivan pienimpiä yhdysosia, vaan osat **riksdag** ja **hus**.

Yhdyssanojen osittamisen ja eliminoimisen tarkkuutta parantava vaikutus tulee puolestaan hyvin esiin hakuaiheessa 148. Hakuaihe sisältää yhdyssanat **ozonskikt** ja **miljöförorening**, jotka ovat molemmat kompositionaalisia yhdyssanoja ja jotka kannattaa jakaa osiin kyselyissä. Samanlainen on tarkkuuksiltaan hakuaihe183, jossa yhdyssanan **dinosaurielämning** osittaminen parantaa tarkkuutta molempien käsittelymenetelmien osalta. Hakuaiheiden joukosta löytyy kuitenkin myös tapauksia, joissa kompositionaalisen yhdyssanan osittaminen ei ole hyödyllistä. Esimerkiksi hakuaiheessa 181 yhdyssanan **kärnvapenprov** osittaminen huonontaa tarkkuutta jopa 65 prosenttiyksikköä. CLEF2003-hakuaiheissa on myös esimerkkejä tapauksista, joissa yhdyssanojen osittaminen on tarkkuuksiltaan huonoiten menestyvä menetelmä, kun taas yhdyssanojen osittaminen eliminointia hyödyntäen on yhtä tehokas menetelmä kuin yhdyssanojen käsittelemättä jättäminen. Esimerkiksi hakuaiheessa 196 SWETWOL:n tekemä virhetulkinta vaikuttaa tähän. SWETWOL epäonnistuu sanan **bankerna** perusmuotoistamisessa, ja mukaan tulee oikean perusmuodon **bank** lisäksi myös muoto **bankerna**. Jälkimmäinen muoto puolestaan on osittamisen myötä **bana kerna** ja **ban kerna**, mikä huonontaa eliminoimattoman kyselyn tarkkuutta. Yhdyssanojen luonne voi selittää hakuaihekohtaisia eroja, mutta epäselvää on myös se, kuinka paljon läheisyysoperaattorien käyttö vaikuttaa hakutuloksiin. Sitä selventävätkin rakenteettomilla kyselysarjoilla saadut tulokset.

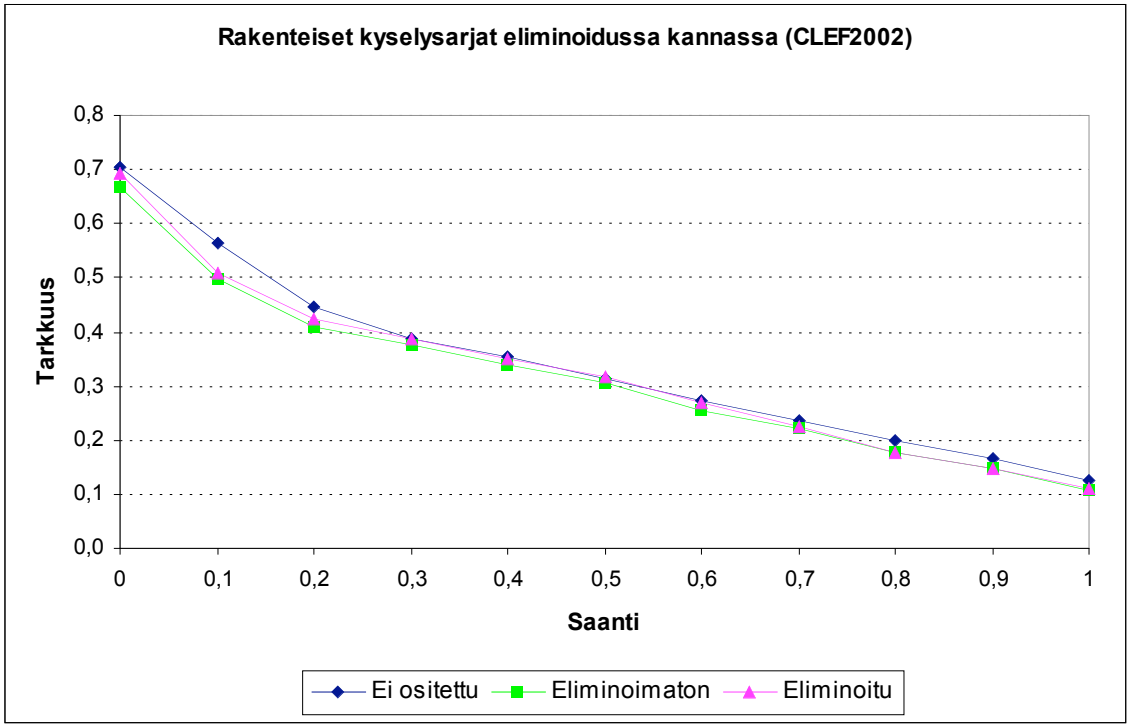
8.1.2 CLEF2002-kyselysarjoilla saadut tulokset

Taulukossa 15 on esitetty keskimääräiset ei-interpoloidut tarkkuusarvot prosentteina CLEF2002-aineiston rakenteisten kyselysarjojen osalta. CLEF2002-aineiston osalta yhdyssanojen osittamisen vaikutus on päinvastainen verrattuna CLEF2003-aineistoon. Eliminoidun kannan osalta perusmuotoisen, yhdyssanoiltaan osittamattoman kyselysarjan tarkkuus on 32,44 prosenttia, kun taas yhdyssanoiltaan ositettu kyselysarja saavuttaa ainoastaan 29,6 prosentin tarkkuuden. Yhdyssanojen osittaminen siis huonontaa kyselyiden tarkkuutta 2,84 prosenttiyksikköä. Myös eliminoimattomassa kannassa ei ositetun ja ositetun kyselyn ero on prosenttiyksiköissä 1,49 prosenttiyksikköä. Tilastollisissa testeissä ei havaittu merkitseviä eroja minkään menetelmän välillä.

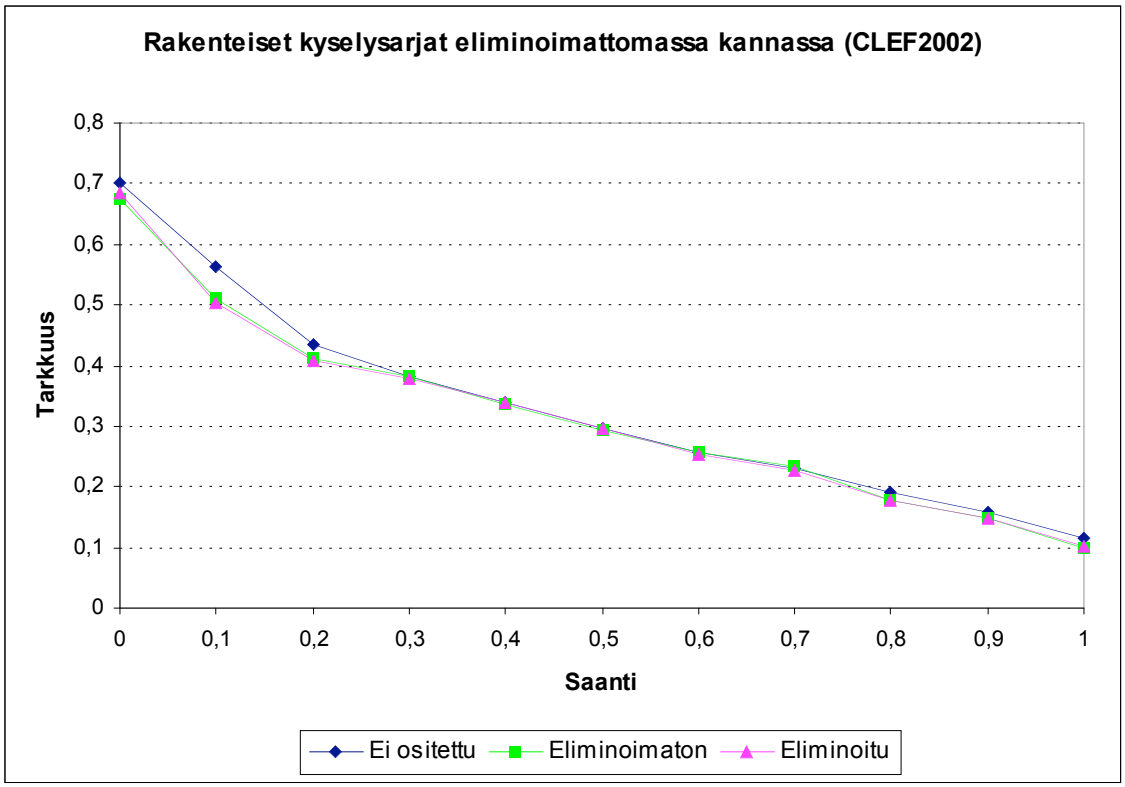
Myöskään yhdyssanojen yhdysosien eliminoiminen ei vaikuta tehokkaalta kyselynmuodostusmenetelmältä. Eliminoidussa kannassa ei ositetun kyselyn tarkkuus on 32,44, kun taas yhdyssanoiltaan ositettu ja eliminoitu kysely saavuttaa 1,8 prosenttiyksikköä huonomman tarkkuuden 30,64 prosenttia. Eliminoimattomassa kannassa perusmuotoinen, yhdyssanoiltaan osittamaton kysely saavuttaa 31,26 prosentin tarkkuuden, siinä missä yhdyssanojen eliminointi huonontaa tarkkuuksia 1,59 prosenttiyksikköä ja tarkkuusarvo on vain 29,67 prosenttia. Näidenkään kyselysarjojen välillä ei havaittu tilastollisesti merkitseviä eroja. Kun katsotaan saanti-tarkkuuskäyrää kuviossa 11, huomataan, että varsinkin alussa yhdyssanoiltaan osittamattoman kyselysarjan tarkkuusarvot ovat eliminoidussa kannassa parempia kuin ne kyselysarjat, joissa yhdyssanoja on käsitelty. Eliminoimattoman kannan osalta menetelmien erot ovat varsin pieniä (katso kuvio 12).

TAULUKKO 15. Keskimääräiset ei-interpoloidut tarkkuusarvot (%) rakenteisten CLEF2002-kyselysarjojen osalta

	ELIMINOITU KANTA	ELIMINOIMATON KANTA
Ei ositettu	32,44	31,26
Ositettu, eliminoimaton	29,60	29,77
Ei ositetun ja ositetun ero (%-yks.)	-2,84	-1,49
Ositettu, eliminoitu	30,64	29,67
Ei ositetun ja eliminoidun ero (%-yks.)	-1,8	-1,59



KUVIO 11. Rakenteisilla CLEF2002-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa.



KUVIO 12. Rakenteisilla CLEF2002-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa.

Myös CLEF2002-kyselysarjojen osalta tutkittiin hakuaihekohtaisia tarkkuusarvoja, ja erot esitetään taulukossa 16. Hakuaiheita on yhteensä 49 kappaletta, ja seitsemässä hakuaiheessa ei ole lainkaan yhdyssanoja. Yhdyssanattomien hakuaiheiden määrä on tosiasiassa suurempi, mutta yhdyssanatulkinnoiksi on laskettu kaikki SWETWOL:n tekemät tulkinnat, myös virheelliset tulkinnat. Verrattuna CLEF2003-kyselysarjoihin tarkkuusarvojen erot eri kyselyiden välillä ovat pienempiä. Monissa hakuaiheissa tarkkuusarvojen parantuminen tai huonontuminen tarkoittaa alle yhden prosenttiyksikön parantumista tai huonontumista. Kun katsotaan eliminoitua kantaa, yhdyssanojen osittaminen ilman eliminointia huonontaa tarkkuutta jopa 23 hakuaiheessa. Ositetun ja eliminoidun kyselyn osalta tarkkuus säilyy useammin samoina, jopa 14 hakuaiheen osalta. Eliminoimattomassa kannassa osittaminen ilman eliminointia lähinnä huonontaa (19 hakuaiheessa) tai parantaa (17 hakuaiheessa) tarkkuusarvoa, kun taas osittaminen eliminointia hyödyntäen huonontaa tarkkuutta 15, parantaa 11 ja pitää tarkkuuden samana 14 hakuaiheessa. Kuten jo mainittiin tarkkuuserot menetelmien välillä ovat kuitenkin varsin pieniä.

TAULUKKO 16. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn (CLEF2002, rakenteiset kyselyt)

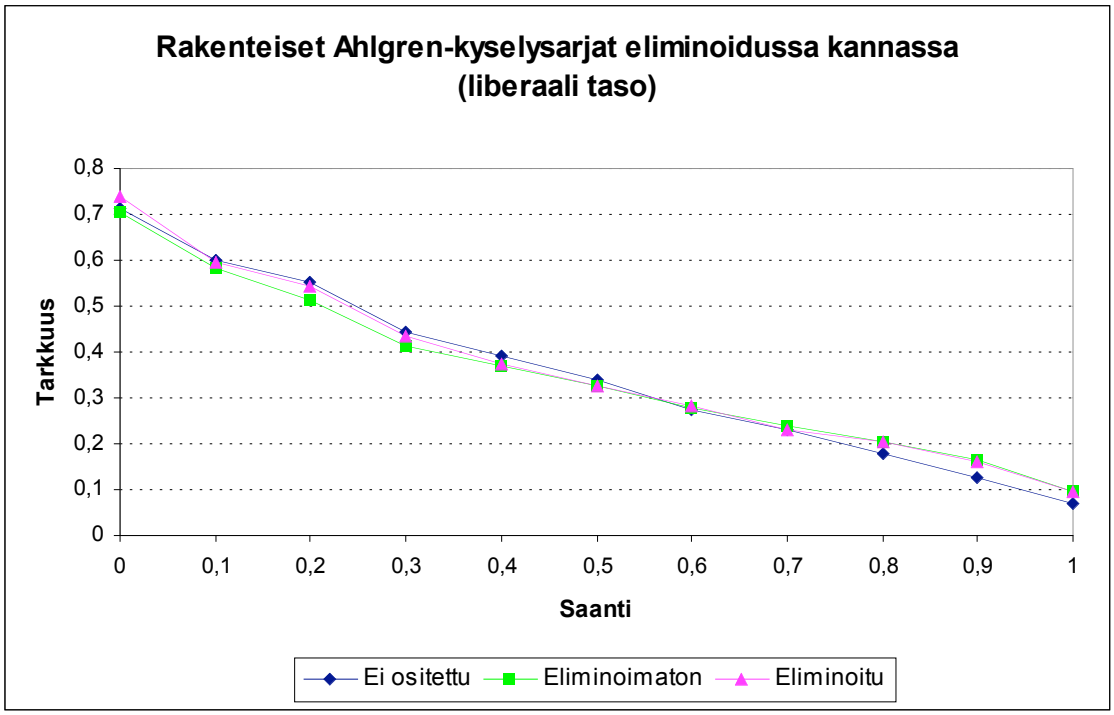
ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	15	11
Huonontunut	23	17
Sama	4	14
Ei yhdyssanoja	7	7
Yhteensä	49	49
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	17	13
Huonontunut	19	15
Sama	6	14
Ei yhdyssanoja	7	7
Yhteensä	49	49

On kuitenkin olemassa joitakin hakuaiheita, joissa tarkkuudet muuttuvat dramaattisemmin suuntaan tai toiseen. Esimerkiksi hakuaiheessa 102 keskimääräinen tarkkuus huonontuu 10 prosenttiyksikön verran yhdyssanojen käsittelemisen myötä. Hakuaihe sisältää kompositionaalisen yhdyssanan **skidtävling**. Kyseessä on esimerkki kompositionaalisesta yhdyssanasta, jonka osittaminen ei ilmeisestikään ole hyödyllistä. Hakuaiheessa 116 keskimääräinen tarkkuus puolestaan parantuu noin 15 prosenttiyksikköä yhdyssanojen **vintersport** ja **guldmedalj** osittamisen myötä. Hakuaiheessa 134 yhdyssanan **rymdsond** osittaminen huonontaa tarkkuutta eliminoidun kannan 41 prosenttiyksikön ja eliminoimattoman kannan 33 prosenttiyksikön verran. Hakuai-

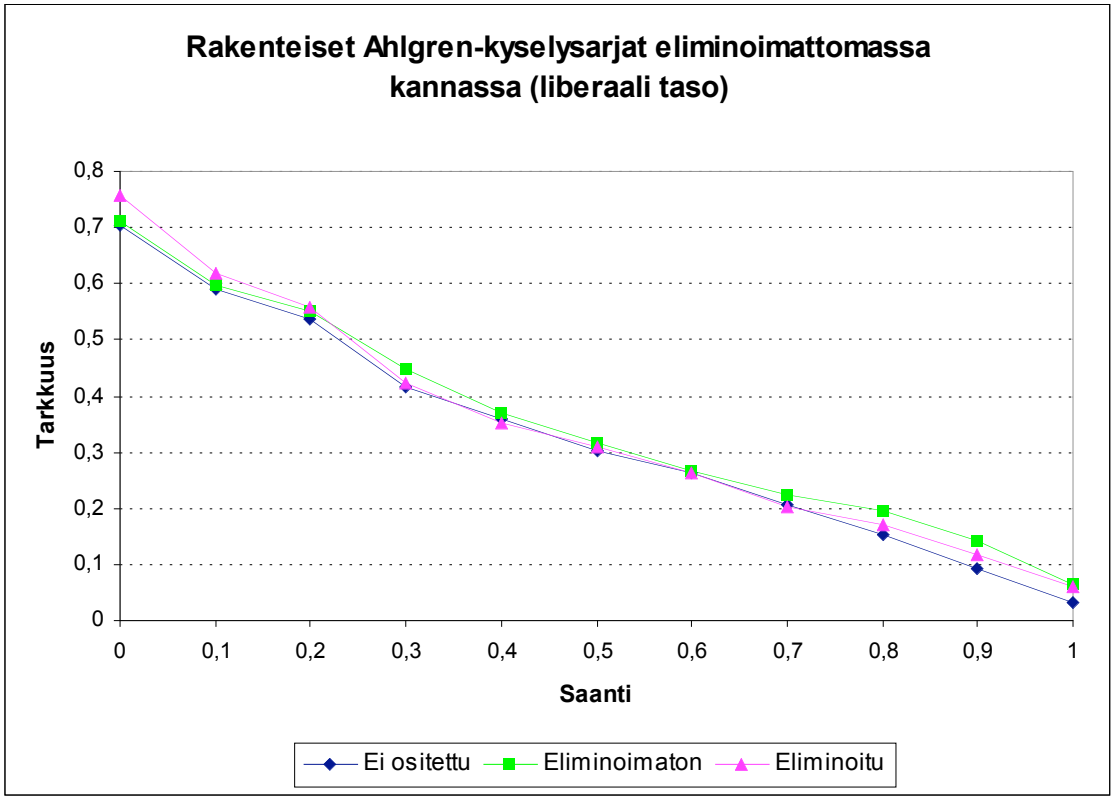
heessa 139 yhdyssanan **fiskekvot** osittaminen puolestaan parantaa keskimääräistä tarkkuutta 14 prosentin verran. Myöskään CLEF2002-kyselyiden osalta ei ole selvää, missä määrin läheisyysoperaattorien käyttö vaikuttaa tuloksiin. CLEF2002-aineisto myös eroaa kahdesta muusta aineistosta pienemmän yhdyssanamääränsä perusteella, millä voi myös olla vaikutusta hakutuloksiin.

8.1.3 Ahlgren-kyselysarjoilla saadut tulokset

Ahlgrenin kokoelmassa on käytössä moniportainen relevanssin arviointi, joten tuloksia arvioidaan kolmella eri tasolla. Kuviossa 13 on esitetty saanti-tarkkuus -käyrä, joka kuvaa Ahlgrenin aineiston kyselysarjoilla saavutettuja tarkkuusarvoja eliminoidussa kannassa liberaalilla relevanssitasolla. Menetelmien väliset erot ovat varsin pieniä. Varsinkin saannin pienemmillä tasoilla eli käyrän alkupäässä ei ositettu kyselysarja ja eliminoitu kyselysarja saavat lähes samankaltaisia tarkkuusarvoja, kun taas eliminoimaton kyselysarja on niitä huonompi. Loppupäässä saannin ollessa 0,7 eliminoimaton kyselysarja ja eliminoitu kyselysarja ohittavat yhdyssanoiltaan osittamattoman kyselysarjan.



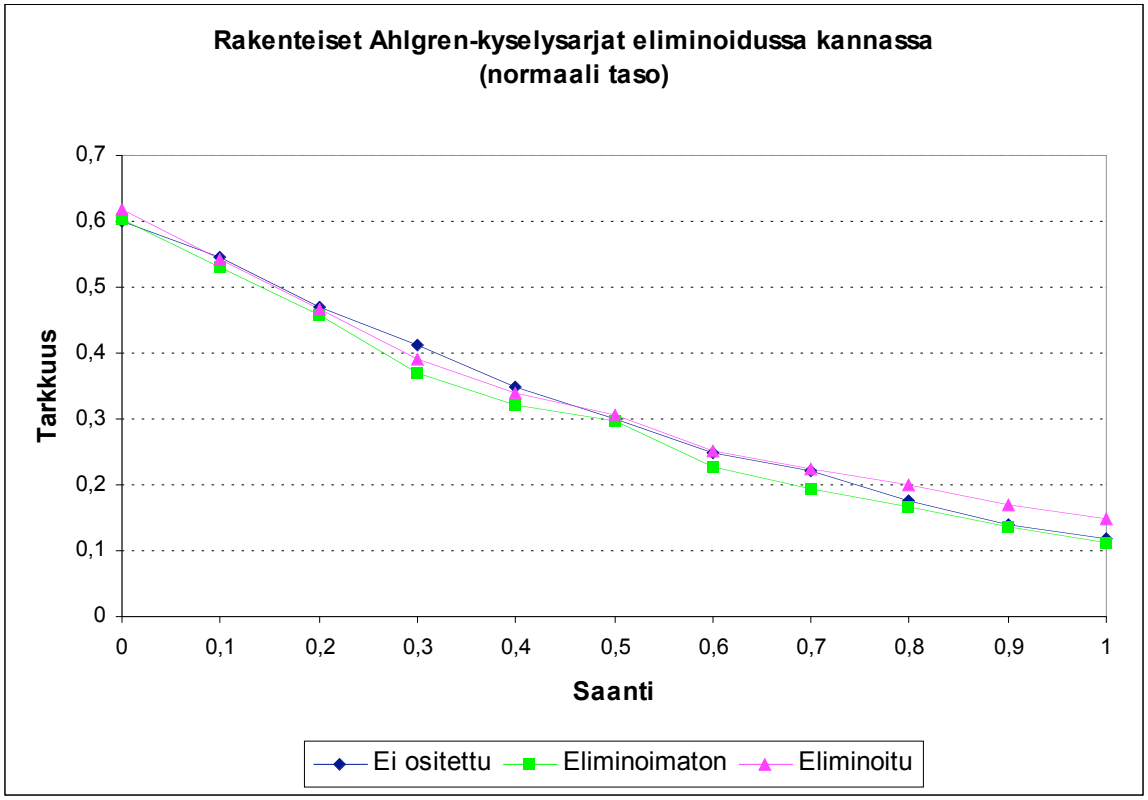
KUVIO 13. Rakenteisilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa liberaalilla relevanssitasolla.



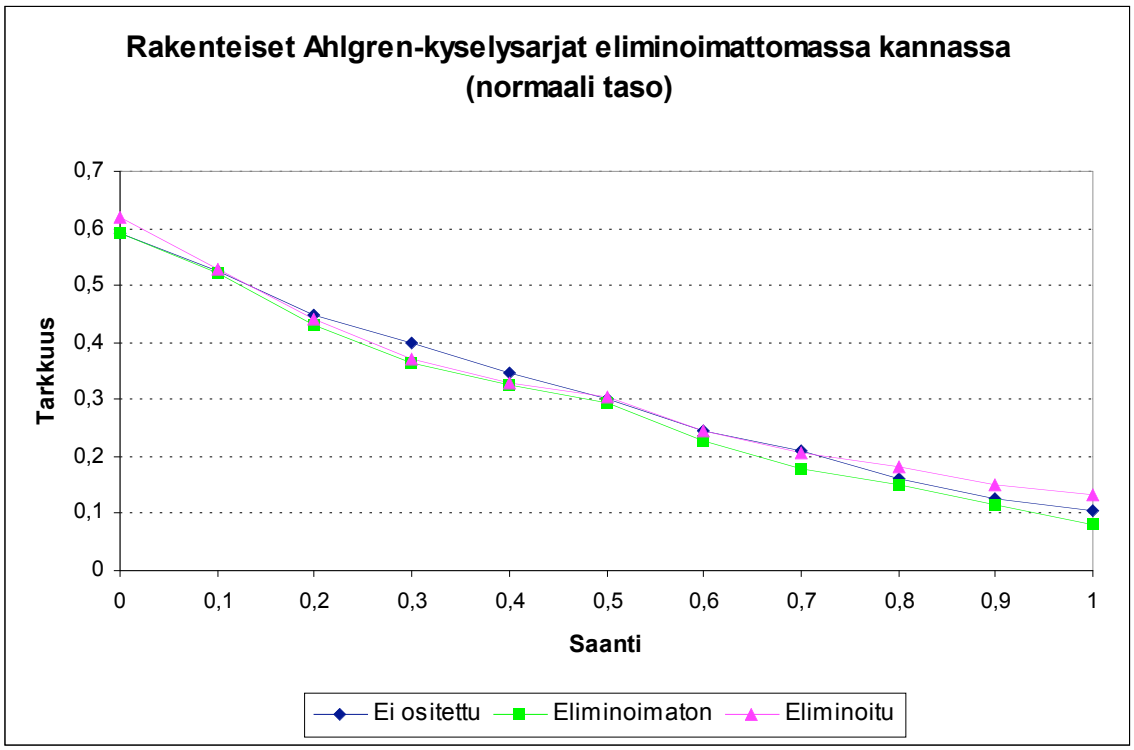
KUVIO 14. Rakenteisilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa liberaalilla relevanssitasolla.

Kuviossa 14 on puolestaan esitetty Ahlgren-kyselysarjoilla saadut tarkkuusarvot eliminoimattomassa kannassa liberaalilla relevanssitasolla. Alussa eliminoitu kyselysarja on paras tarkkuusarvoiltaan, kun taas eliminoimaton ja ei ositettu ovat tarkkuuksiltaan samankaltaisia. Saannin kasvaessa eliminoimaton parantaa asemiaan. Loppupäässä yhdyssanoiltaan käsitellyt kyselysarjat ovat samoin kuin eliminoidussakin kannassa ohittaneet ei ositetun kyselyn. Erot ovat kuitenkin melko pieniä.

Kuviossa 15 on puolestaan esitetty saanti-tarkkuus -käyrä Ahlgrenin kyselysarjojen tarkkuuksista eliminoidussa kannassa normaalilla relevanssitasolla. Menetelmien välillä ei ole suuria eroja, ja varsinkin alussa erot ovat pieniä. Eliminoimaton kyselysarja on tarkkuusarvoiltaan huonoin menetelmä. Saannin kasvaessa eliminoitu kyselysarja ohittaa ei ositetun kyselysarjan.



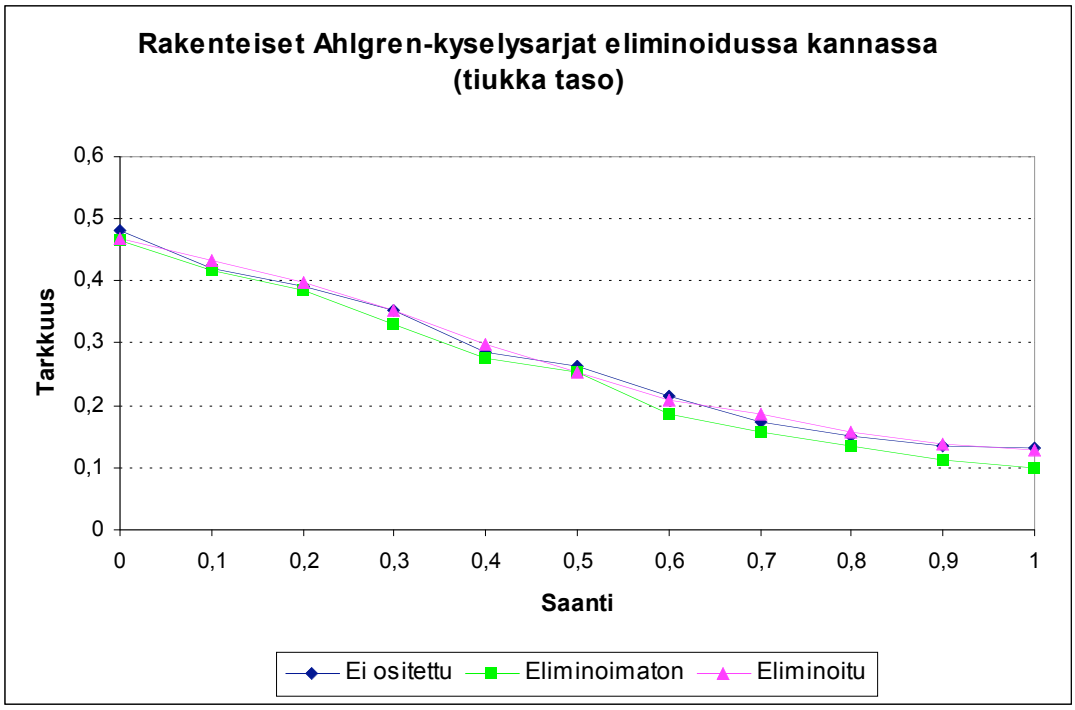
KUVIO 15. Rakenteisilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa normaalilla relevanssitasolla.



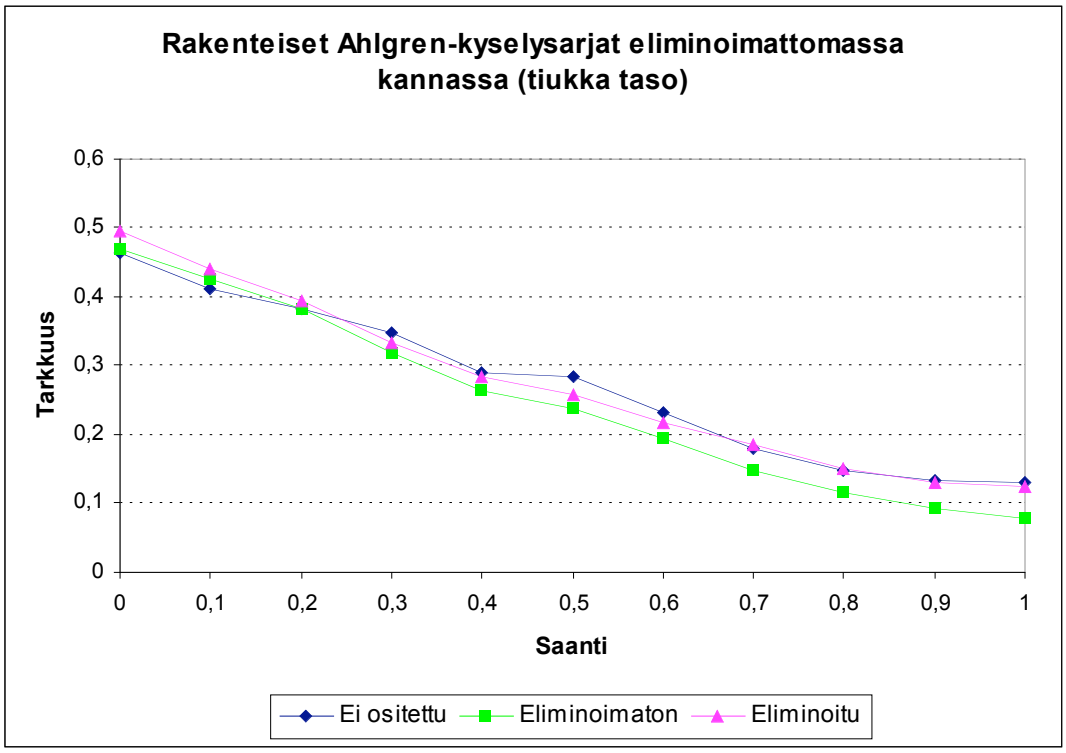
KUVIO 16. Rakenteisilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa normaalilla relevanssitasolla.

Kuviossa 16 on esitetty Ahlgren-kyselysarjoilla saadut tarkkuusarvot eliminoimattomassa kannassa normaalilla relevanssitasolla. Kuvio muistuttaa paljon kuviota 15, joka kuvaa eliminoidun kannan tuloksia. Eliminoimaton kyselysarja on tässäkin huonoiten menestyvä menetelmä. Myös eliminoimattomassa kannassa eliminoitu kyselysarja parantaa saannin kasvaessa asemiaan. Erot menetelmien välillä ovat kuitenkin pieniä.

Kuviossa 17 on esitetty saanti-tarkkuus -käyrä Ahlgren-kyselysarjoilla saaduista tarkkuuksista eliminoidussa kannassa tiukalla relevanssitasolla. Tässäkään kuviossa vertailtujen kyselysarjojen välillä ei ole suuria eroja. Varsinkin saannin kasvaessa eliminoimaton kyselysarja on menetelmänä huonompi kuin kaksi muuta kyselysarjaa, jotka ovat tarkkuuksiltaan melko samankaltaisia. Kuviossa 18 puolestaan on kuvattu Ahlgren-kyselysarjojen tarkkuusarvot eliminoimattomassa kannassa tiukalla relevanssitasolla. Kuvio muistuttaa kuviota 17. Varsinkin alussa menetelmien väliset erot ovat pieniä. Saannin kasvaessa eliminoimaton kyselysarja on selvästi huonompi kuin kaksi muuta menetelmää, jotka ovat tarkkuuksiltaan samankaltaisia. Saannin ollessa 0,5 ei ositettu kyselysarja ohittaa tarkkuusarvoissa kaksi muuta kyselysarjaa, mutta loppua kohden ei ositettu ja eliminoitu ovat tasaväkisiä.



KUVIO 17. Rakenteisilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa tiukalla relevanssitasolla.



KUVIO 18. Rakenteisilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa tiukalla relevanssitasolla.

Taulukossa 17 on kooste rakenteisten Ahlgren-kyselysarjojen tarkkuuksista kolmella eri relevanssitasolla. Kun katsotaan prosenttiyksiköittäin menetelmien välisiä eroja eliminoidun kannan osalta, huomataan, että yhdyssanojen osittaminen huonontaa kyselysarjojen tarkkuusarvoa mitä tiukemmalle relevanssitasolle siirrytään. Liberaalilla relevanssitasolla huononeminen on vain 0,1 prosenttiyksikköä, kun taas tiukalla relevanssitasolla jopa 1,43 prosenttiyksikköä. Yhdyssanojen eliminoiminen parantaa tarkkuutta vain varsin vähän vaihdellen liberaalin tason 0,69 prosenttiyksikön parannuksesta tiukan tason 0,4 prosenttiyksikköön. Myös eliminoimattomassa kannassa yhdyssanojen yhdysosien eliminointi parantaa tarkkuutta verrattuna yhdyssanoiltaan osittamattomaan kyselysarjaan. Parannus on prosenttiyksiköissä ilmaistuna suurinta liberaalilla relevanssitasolla ja tiukimmalla relevanssitasolla enää ainoastaan 0,2 prosenttiyksikköä. Selvä poikkeama taulukossa on yhdyssanoiltaan ositettu, mutta eliminoimaton kyselysarja eliminoimattomassa kannassa. Liberaalilla relevanssitasolla yhdyssanojen osittaminen parantaa tarkkuusarvoa jopa 2,66 prosenttia, mutta siirryttäessä tiukemmille relevanssitasoille tarkkuudet alkavat huonontua siten, että normaalilla relevanssitasolla osittaminen huonontaa tarkkuutta 1,38 prosenttiyksikköä ja tiukalla relevanssitasolla jopa 2,51 prosenttiyksikköä. Tilastollisissa testeissä ei kuitenkaan havaittu merkitseviä eroja.

Jos lähestytään asiaa siitä näkökulmasta, että halutaan löytää vain erittäin relevantteja dokumentteja (tiukka relevanssitaso), yhdyssanojen käsittelystä ei ole suurta hyötyä. Tiukalla relevanssitasolla yhdyssanojen osittaminen huonontaa molemmissa kannoissa tarkkuusarvoja (1,43 vastaan 2,51 prosenttiyksikköä). Yhdyssanojen yhdysosien eliminoiminen puolestaan ei eroa paljoakaan vertailukohdasta eli perusmuotoisesta kyselystä, jonka yhdyssanoja ei ole lainkaan käsitelty. Prosenttiyksiköissä erot ovat ainoastaan eliminoidussa kannassa 0,4 ja eliminoimattomassa kannassa 0,2 prosenttiyksikköä. Nämäkään erot eivät ole tilastollisesti merkitseviä.

TAULUKKO 17. Keskimääräiset ei-interpoloidut tarkkuusarvot (%) rakenteisten Ahlgren-kyselysarjojen osalta.

ELIMINOITU KANTA, kyselyt	Liberaali taso	Normaali taso	Tiukka taso
Ei ositettu	33,75	30,59	25,73
Ositettu, ei eliminoitu	33,65	29,21	24,3
Ei ositetun ja ositetun ero (%-yks.)	-0,1	-1,38	-1,43
Ositettu, eliminoitu	34,44	31,36	26,13
Ei ositetun ja eliminoidun ero (%-yks.)	0,69	0,77	0,4
ELIMINOIMATON KANTA, kyselyt	**Liberaali taso**	**Normaali taso**	**Tiukka taso**
Ei ositettu	31,25	29,43	25,93
Ositettu, ei eliminoitu	33,91	28,05	23,42
Ei ositetun ja ositetun ero (%-yks.)	2,66	-1,38	-2,51
Ositettu, eliminoitu	32,75	30,15	26,13
Ei ositetun ja eliminoidun ero (%-yks.)	1,5	0,72	0,2

Myös Ahlgren-kyselysarjojen osalta tutkittiin hakuaihekohtaisia tarkkuusarvoja, ja tarkkuuksien muutokset on esitetty liberaalin relevanssitason osalta taulukossa 18. Liberaalilla relevanssitasolla tulokset saatiin kullekin 51 hakuaiheen kyselylle. Kolmessa hakuaiheessa ei ole lainkaan yhdyssanoja, ja saaduissa tuloksissa ei ole eroja eri kyselyjen suhteen. Sekä eliminoidussa että eliminoimattomassa kannassa yhdyssanoiltaan ositettu ja eliminoimaton kyselysarja on tarkkuusarvoiltaan usein joko parantunut (21 ja 20 hakuaiheessa) tai huonontunut (20 ja 23 hakuaiheessa). Yhdyssanojen osittaminen siis muuttaa voimakkaasti tarkkuusarvoja joko suuntaan tai toiseen. Ainoastaan eliminoidussa kannassa seitsemän tai eliminoimattomassa kannassa viiden hakuaiheen osalta tarkkuusarvot säilyvät samanlaisina verrattuna yhdyssanoiltaan osittamattomaan kyselysarjaan. Kun katsotaan yhdyssanoiltaan ositettuja ja eliminoituja kyselyitä, tarkkuusarvot säilyvät molemmissa kannoissa monissa hakuaiheissa samoina (22 ja 22).

Eliminoidussa kannassa 13 hakuaiheessa tapahtuu tarkkuusarvojen huonontumista ja 13 parantumista yhdyssanojen osittamisen ja yhdysosien eliminoimisen myötä. Eliminoimattoman kannan osalta vastaavat luvut ovat 14 ja 12.

TAULUKKO 18. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn, rakenteiset Ahlgren-kyselyt, liberaali taso

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	21	13
Huonontunut	20	13
Sama	7	22
Ei yhdyssanoja	3	3
Yhteensä	51	51
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	20	14
Huonontunut	23	12
Sama	5	22
Ei yhdyssanoja	3	3
Yhteensä	51	51

TAULUKKO 19. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn, rakenteiset Ahlgren-kyselyt, normaali taso

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	19	16
Huonontunut	18	9
Sama	10	22
Ei yhdyssanoja	3	3
Yhteensä	50	50
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	17	13
Huonontunut	24	13
Sama	6	21
Ei yhdyssanoja	3	3
Yhteensä	50	50

Taulukossa 19 on esitetty samat tarkkuusarvojen muutokset normaalin relevanssitason osalta. Normaalilla relevanssitasolla tulokset saatiin 51 hakuaiheesta 50 hakuaiheen kyselyille. Kolmessa hakuaiheessa ei ole lainkaan yhdyssanoja, ja saaduissa tuloksissa ei ole eroja eri kyselyjen suhteen. Tarkkuuden muutokset ovat normaalilla tasolla samankaltaisia kuin liberaalilla tasolla. Eliminoidussa kannassa ositettu, ei eliminoitu kyselysarja aiheuttaa tarkkuuden huonontumista 18 ja parantumista 19 hakuaiheen osalta, kun taas 10 hakuaiheessa tarkkuudet säilyvät samoina. Ositettu, eliminoitu kyselysarja puolestaan muuttaa edellistä vähemmän tarkkuuksia suuntaan tai

toiseen. Jopa 22 hakuaiheen osalta tarkkuus pysyy samana, 16 hakuaiheessa parantuu ja 9 hakuaiheen osalta huonontuu. Eliminoimattoman kannan osalta muutokset ovat samankaltaisia. Yhdyssanojen osittaminen aiheuttaa tarkkuuden vaihteluita (parantuminen 17, huonontuminen 24). Samana tarkkuus pysyy ainoastaan kuudessa hakuaiheessa. Yhdyssanojen osittaminen eliminointia hyödyntäen puolestaan pitää tarkkuuden samana jopa 21 hakuaiheessa.

TAULUKKO 20. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn, rakenteiset Ahlgren-kyselyt, tiukka taso

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	16	13
Huonontunut	14	6
Sama	11	22
Ei yhdyssanoja	3	3
Yhteensä	44	44
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	16	11
Huonontunut	15	8
Sama	10	22
Ei yhdyssanoja	3	3
Yhteensä	44	44

Taulukossa 20 on esitetty hakuaiheiden tarkkuusarvojen muutokset tiukalla relevanssitasolla. Tiukalla relevanssitasolla tulokset saatiin 51 hakuaiheesta 44 hakuaiheen kyselyille. Kolmessa hakuaiheessa ei ole lainkaan yhdyssanoja, ja saaduissa tuloksissa ei ole eroja eri kyselyjen suhteen. Myös tiukalla relevanssitasolla tarkkuuden muutokset ovat edellä esitetyn kaltaisia. Sekä eliminoidussa että eliminoimattomassa kannassa yhdyssanojen osittaminen aiheuttaa suurempia heilahteluita tarkkuuden suhteen, kun taas yhdyssanojen eliminointi pitää tarkkuuden useammin samana.

Hakuaihekohtaisten tarkkuuksien tutkiminen osoittaa, että yhdyssanojen osittaminen aiheuttaa suurempia tarkkuuden heilahteluita suuntaan tai toiseen kuin yhdyssanojen eliminoiminen. Myös keskimääräisiä tarkkuuksia koskevissa hakutuloksissa (taulukko 17) yhdyssanojen osittaminen (ilman eliminointia) aiheuttaa suurempia eroja tarkkuuksien välille kuin yhdyssanojen osittaminen eliminointia hyödyntäen.

Kun tutkitaan hakuaiheiden sisältämiä yhdyssanoja, huomataan, että hakuaiheessa 5 yhdyssanan **medlemsland** osittaminen parantaa tarkkuutta kaikilla relevanssitasoilla. Myös hakuaiheessa 22

yhdyssanojen **flygplansolycka** ja **landningsbana** osittaminen parantaa tarkkuutta kaikilla relevanssitasoilla. Samanlainen esimerkki on hakuaiheen 26 yhdyssana **vindkraft**, jonka osittaminen parantaa tarkkuutta. Hakuaiheessa 57 tarkkuus puolestaan huononee yhdyssanojen **rättsprocess**, **frankrike** ja **domstol** osittamisen myötä. Joissakin tapauksissa menetelmän menestyminen yllättää. Esimerkiksi voisi kuvitella, että SWETWOL:n väärä tulkinta sanasta **styrkor** (**styra kor**) hakuaiheessa 48 huonontaisi tarkkuutta, mutta ositettu ja ei eliminoitu kysely suoriutuukin paremmin tarkkuusvertailussa kuin ositettu ja eliminoitu kysely, jossa yhdysosia ei esiinny. Samoin hakuaiheessa 61 voisi kuvitella, että yhdyssanan **oljeledning** osittaminen parantaisi tarkkuutta, mutta vaikutus onkin täysin päinvastainen. Joitakin tarkkuuksien välillä olevia eroja on vaikea selittää, koska kyselyt ovat varsin samankaltaisia. Näihin eroihin vaikuttavatkin varmasti yhdyssanojen luonnetta enemmän hakemistot ja dokumenttien sisältämät sanat. Esimerkiksi hakuaiheessa 62 tarkkuudessa on eroja ei eliminoidun ja eliminoidun kyselyn välillä, vaikka hakuaiheesta muodostetut kyselyt ovat sanoiltaan samankaltaiset. Läheisyysoperaattorien käyttäminen kyselyissä voi myös vaikuttaa tarkkuuksiin.

Yhteenvetona voidaan todeta, että rakenteisten kyselysarjojen vertailussa on käytetty kolmea eri aineistoa ja saadut tulokset ovat ristiriitaisia. Kaiken kaikkiaan eri kyselysarjojen väliset erot ovat varsin pieniä, eikä tilastollisesti merkitseviä eroja havaittu. CLEF2003-aineisto antaa kuitenkin viitteitä siitä, että yhdyssanoja kannattaisi käsitellä ruotsinkielisessä tiedonhaussa. Eliminoidussa kannassa tarkkuusarvot ovat sekä eliminoimattoman että eliminoidun kyselysarjan osalta parempia kuin yhdyssanoiltaan osittamattoman kyselysarjan tarkkuudet. Eliminoimattomassa kannassa yhdyssanojen osittaminen parantaa tarkkuuksia, kun taas yhdyssanojen eliminoiminen ei eroa kovinkaan suuresti osittamattomasta kyselysarjasta.

CLEF2002-aineisto antaa puolestaan täysin päinvastaisia tuloksia, ja tämän aineiston osalta sekä yhdyssanojen osittaminen että eliminoiminen huonontaa keskimääräistä tarkkuusarvoa verrattuna perusmuotoiseen, osittamattomaan kyselysarjaan. CLEF2002-aineiston osalta huomioitavaa on se, että aineisto eroaa kahdesta muusta pienemmän yhdyssanamäärän perusteella. Ahlgren-aineisto puolestaan antaa viitteitä siitä, että yhdyssanojen osittaminen ilman eliminointia ei ole kannattava menetelmä. Eliminoidussa kannassa yhdyssanojen osittaminen ilman eliminointia huonontaa tarkkuusarvoja kaikilla relevanssitasoilla. Eliminoimattomassa kannassa yhdyssanojen osittaminen parantaa tarkkuutta liberaalilla relevanssitasolla, mutta muilla relevanssitasoilla tarkkuusarvot huononevat. Yhdyssanojen osittaminen ja eliminoiminen näyttäisi puolestaan parantavan hiukan kyselysarjojen tarkkuusarvoja. Erot menetelmien välillä eivät kuitenkaan ole

suuria eivätkä tilastollisesti merkitseviä. Seuraavassa luvussa nähdään rakenteettomilla kyselysarjoilla saadut tulokset, ja voidaan vertailla esimerkiksi sitä, millainen on läheisyysoperaattorien vaikutus hakutuloksiin.

8.2 Rakenteettomat kyselysarjat

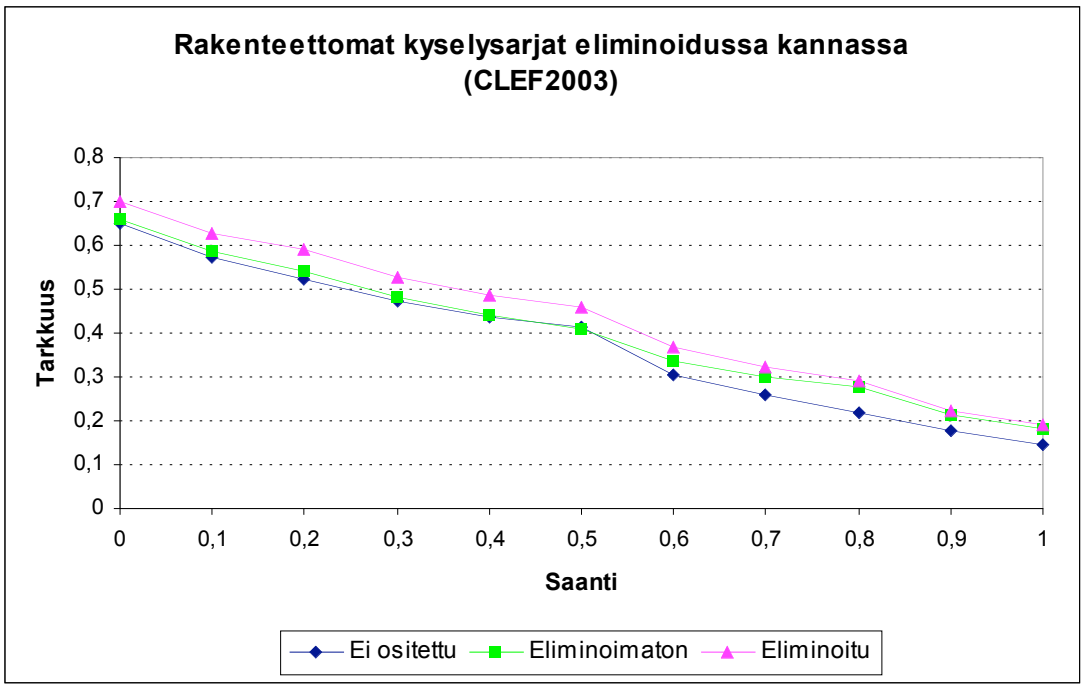
Seuraavaksi esittelen eri rakenteettomilla #combine-operaattorista ja sanalistasta muodostuvilla kyselysarjoilla saadut tulokset kolmen aineiston osalta.

8.2.1 CLEF2003-kyselysarjoilla saadut tulokset

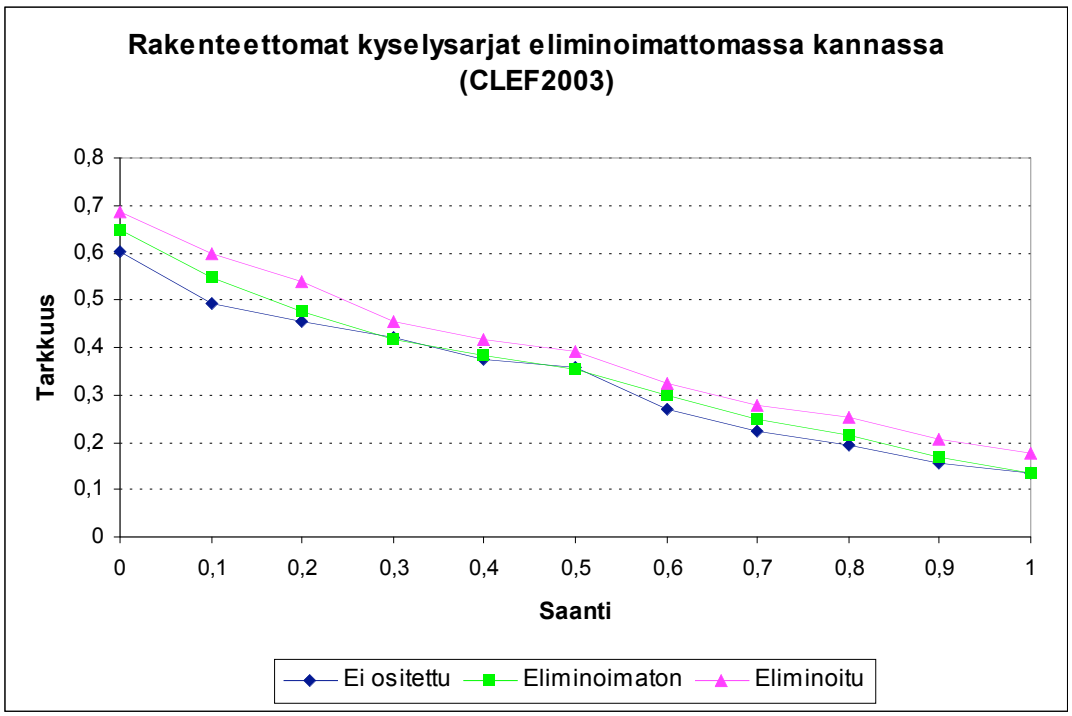
TAULUKKO 21. Keskimääräiset ei-interpoloidut tarkkuusarvot (%) rakenteettomien CLEF2003-kyselysarjojen osalta

	ELIMINOITU KANTA	ELIMINOIMATON KANTA
Ei ositettu	36,25	31,87
Ositettu, eliminoimaton	38,84	33,77
Ei ositetun ja ositetun ero (%-yks.)	2,59	1,9
Ositettu, eliminoitu	42,04	37,48
Ei ositetun ja eliminoidun ero (%-yks.)	5,79	5,61

Taulukossa 21 on esitetty CLEF2003-kyselysarjojen rakenteettomien versioiden tulokset. Verrattuna rakenteisiin kyselyihin tarkkuusarvot ovat korkeampia. Kun katsotaan eliminoitua kantaa, huomataan, että molemmat yhdyssanojen käsittelytavat parantavat tarkkuusarvoja. Yhdyssanojen osittaminen ilman eliminointia parantaa keskimääräistä tarkkuutta osittamattomaan kyselysarjaan verrattuna 2,59 prosenttiyksikköä. Yhdyssanojen osittaminen eliminointia hyödyntäen parantaa tarkkuutta osittamattomaan kyselysarjaan verrattuna jopa 5,79 prosenttiyksikköä. Eliminoimattomassa kannassa saadut tulokset ovat samansuuntaisia. Yhdyssanojen osittaminen eliminointia hyödyntämättä parantaa tarkkuutta osittamattomaan kyselysarjaan verrattuna 1,9 prosenttiyksikköä. Yhdyssanojen osittaminen ja eliminoiminen puolestaan parantaa tarkkuutta jopa 5,61 prosenttiyksikköä. Myös kuviosta 19 ja 20 huomataan, että yhdyssanojen osittaminen eliminointia hyödyntäen on parhaiten menestyvä menetelmä sekä eliminoidussa että eliminoimattomassa kannassa. Rakenteettomien kyselysarjojen sisällä menetelmien väliset erot eivät kuitenkaan ole tilastollisesti merkitseviä. Sen sijaan kun verrataan rakenteettomien kyselysarjojen tarkkuuksia rakenteisten kyselysarjojen tarkkuuksiin, havaittiin tilastollisesti merkitsevä ero ($p<0,05$) rakenteisten ja rakenteettomien eliminoitujen kyselyjen tarkkuuksien välillä eliminoidussa kannassa.



KUVIO 19. Rakenteettomilla CLEF2003-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa.



KUVIO 20. Rakenteettomilla CLEF2003-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa.

Taulukossa 22 on esitetty hakuaihekohtaiset tarkkuuksien vaihtelut CLEF2003-aineiston rakenteettomien kyselyiden osalta. Yhteensä 54 hakuaiheesta kuudessa ei ole lainkaan yhdyssanoja, joten saadut tulokset eivät eroa tarkkuuksiltaan. Rakenteisiin kyselyihin verrattuna tarkkuusarvoista suurempi osa paranee yhdyssanojen käsittelyn myötä. Molemmissa kannoissa ja molemmilla menetelmillä esiintyvät enemmistössä ne hakuaiheet, joissa tarkkuusarvo on parantunut yhdyssanojen käsittelyn myötä. Yhdyssanoiltaan ositetut ja eliminoidut kyselyt parantavat hiukan enemmän tarkkuutta, 32 ja 31 hakuaiheessa verrattuna toisen yhdyssanamenetelmän 28 ja 26 hakuaiheeseen. Ositettujen ja ei eliminoitujen kyselyjen joukossa on myös enemmän sellaisia hakuaiheita, joissa tarkkuus huononee. Eliminoidussa kannassa tällaisia on 15 ja eliminoimattomassa kannassa 19 hakuaihetta. Tämä viittaisi siihen, että CLEF2003-aineiston osalta paras menetelmä on myös hakuaihekohtaisten tarkkuuksien osalta yhdyssanojen osittaminen eliminointia hyödyntäen.

TAULUKKO 22. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn (CLEF2003, rakenteettomat kyselyt)

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	28	32
Huonontunut	15	8
Sama	5	8
Ei yhdyssanoja	6	6
Yhteensä	54	54
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	26	31
Huonontunut	19	9
Sama	3	8
Ei yhdyssanoja	6	6
Yhteensä	54	54

Kun tutkitaan tarkemmin yhdyssanoja niistä hakuaiheista, joissa tarkkuuden muutokset ovat suuria, huomataan, että myös rakenteettomien kyselyiden osalta löytyy esimerkkejä hakuaiheista, joissa kompositionaalisen yhdyssanan osittaminen ei kannata. Esimerkiksi hakuaiheessa 163 yhdyssanojen **regelverk** ja **lagstiftning** osittaminen huonontaa tarkkuutta. Näin käy myös hakuaiheessa 175 yhdyssanojen **miljöskada** ja **sockerindustri** osalta. Hakuaiheiden rakenteettomista kyselyistä löytyy kuitenkin myös paljon esimerkkejä kompositionaalisista yhdyssanoista, joiden osittaminen kannattaa, esimerkiksi hakuaiheissa 144 (**diamantindustri**), 161 (**dietproblem**, **glutenallergiker**) ja 174 (**krucifixstrid**). Lisäksi rakenteettomissa kyselyissä on monia esimerkkejä virhetulkinnoista, joiden vuoksi tarkkuus laskee erityisesti eliminoimattomissa kyselyissä. Esi-

merkiksi tämän aiheuttaa hakuaiheessa 180 sanan **konkurs** väärä osiin jakaminen ja kyselyssä 187 sanojen **rapporter** ja **transport** virheellinen osittaminen. Kun verrataan yhdyssanaesimerkkejä rakenteettomissa ja rakenteisissa kyselyissä, ainoastaan hakuaiheet 183 ja 186 esiintyvät molemmissa eniten tarkkuuden muutoksia aiheuttaneiden hakuaiheiden listalla. Rakenteisten kyselyjen tavoin myös rakenteettomissa kyselyissä hakuaiheen 183 **dinosaurielämning** on tarkoituksenmukaista osittaa. Näin on myös hakuaiheen 186 yhdyssanojen **regeringskoalition** ja **purpurkoalition** suhteen.

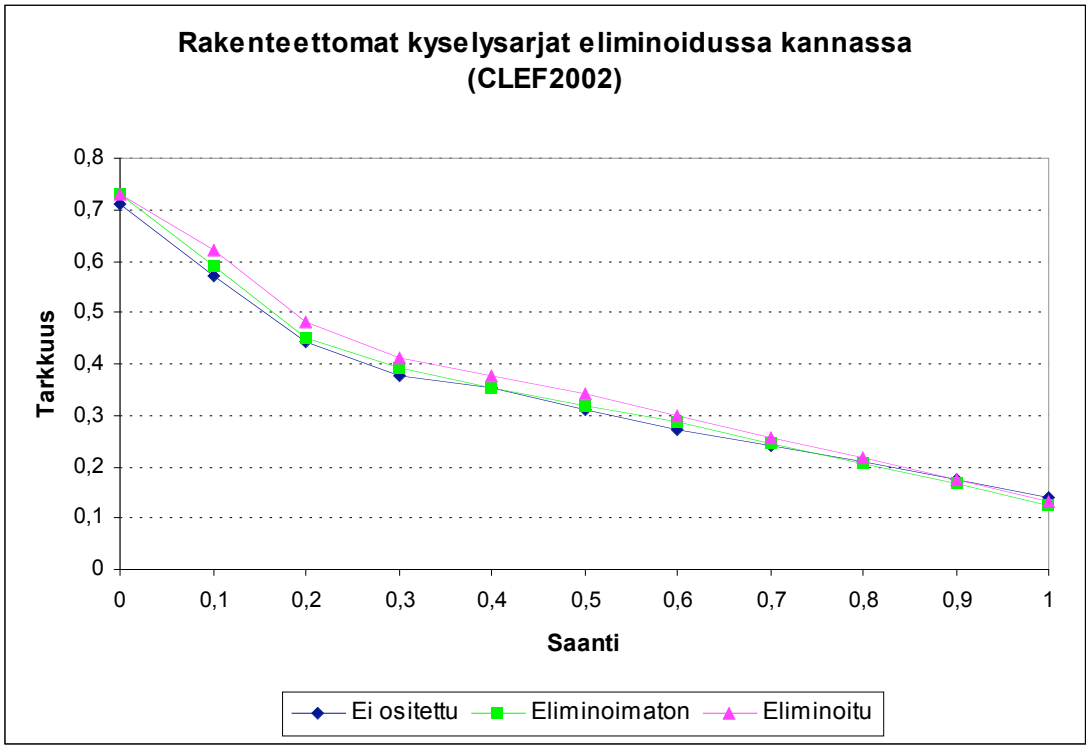
8.2.2 CLEF2002-kyselysarjoilla saadut tulokset

TAULUKKO 23. Keskimääräiset ei-interpoloidut tarkkuusarvot (%) rakenteettomien CLEF2002-kyselysarjojen osalta

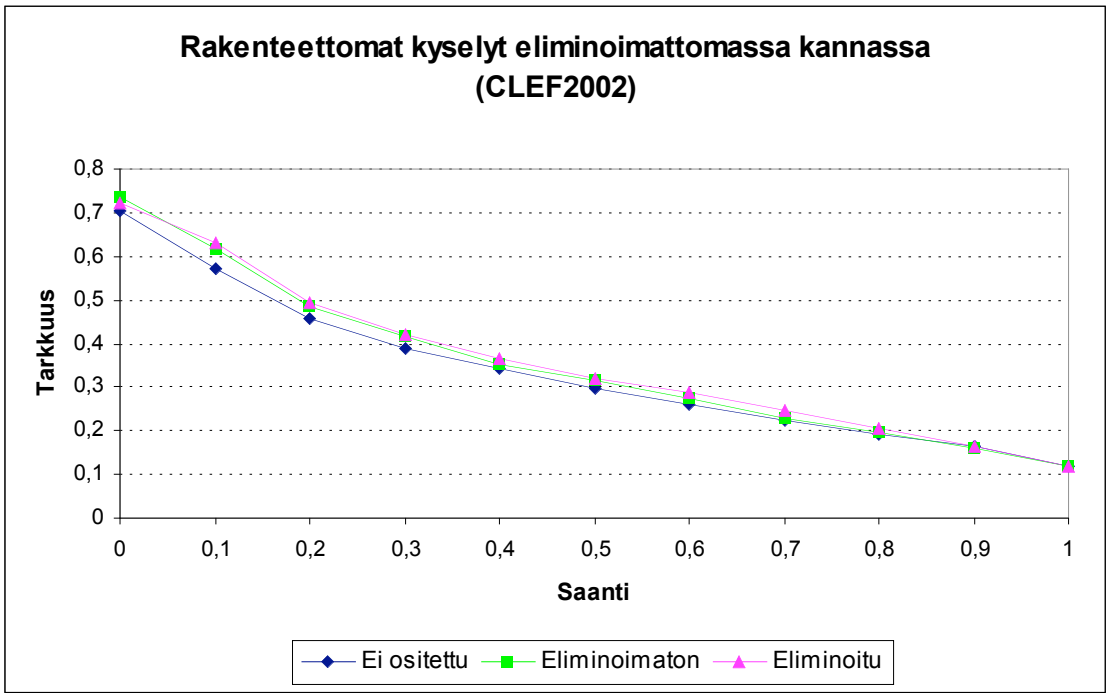
	ELIMINOITU KANTA	ELIMINOIMATON KANTA
Ei ositettu	32,63	31,76
Ositettu, eliminoimaton	32,94	33,29
Ei ositetun ja ositetun ero (%-yks.)	0,31	1,53
Ositettu, eliminoitu	34,82	34,15
Ei ositetun ja eliminoidun ero (%-yks.)	2,19	2,39

Taulukossa 23 on esitetty tarkkuuksien keskiarvot, jotka on saatu rakenteettomilla CLEF2002-kyselysarjoilla. Myös CLEF2002-aineiston osalta yhdyssanojen käsittely rakenteettomissa kyselyissä parantaa hakujen tarkkuuksia verrattuna kyselysarjaan, jossa yhdyssanoja ei ole käsitelty. Eliminoidussa kannassa ei ositetun kyselysarjan ja ositetun, eliminoimattoman kyselysarjan välillä oleva ero on tosin varsin pieni, vain 0,31 prosenttiyksikköä. Ositetun ja eliminoidun kyselysarjan tarkkuus on puolestaan 2,19 prosenttiyksikköä parempi kuin osittamattoman kyselysarjan tarkkuus. Eliminoimattomassa kannassa yhdyssanojen osittaminen eliminointia hyödyntämättä parantaa tarkkuutta 1,53 prosenttiyksikköä verrattuna osittamattomaan kyselysarjaan. Eliminointi puolestaan parantaa tarkkuutta 2,39 prosenttiyksikköä. Erot menetelmien välillä ovat pienempiä kuin CLEF2003-aineistossa, mikä käy myös ilmi kuvioista 21 ja 22. Tilastollisissa testeissä ei myöskään havaittu merkitseviä eroja rakenteettomien kyselysarjojen tuloksissa. Kun verrataan rakenteettomilla kyselysarjoilla saatuja tuloksia rakenteisten kyselysarjojen tuloksiin, huomataan, että tulokset ovat päinvastaisia. Rakenteisten kyselysarjojen osalta molemmat yhdyssanojen käsittelymenetelmät huonontavat tarkkuusarvoja molemmissa kannoissa, kun taas rakenteettomien kyselysarjojen osalta vaikutukset ovat päinvastaiset. Tilastollisissa testeissä havaittiinkin merkitsevä ero ($p<0,05$) rakenteisten ja rakenteettomien eliminoitujen kyselyjen tarkkuuksien

välillä eliminoimattomassa kannassa. Mielenkiintoista on se, millainen rooli läheisyysoperaattorien käytöllä on tuloksissa.



KUVIO 21. Rakenteettomilla CLEF2002-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa.



KUVIO 22. Rakenteettomilla CLEF2002-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa.

Taulukossa 24 on esitetty hakuaihekohtaisten tarkkuuksien vaihtelut rakenteettomien CLEF2002-kyselysarjojen osalta. Yhteensä 49 hakuaiheesta seitsemässä ei ole lainkaan yhdyssanoja. Rakenteettomien CLEF2003-kyselysarjojen vastaaviin arvoihin verrattuna erot ovat tasaisempia, ja hakuaihekohtaisissa tarkkuuksissa tapahtuu tasaisesti muutoksia sekä huonompaan että parempaan suuntaan. Erityisesti näin on ositettujen, ei eliminoitujen kyselyjen kohdalla. Ositettujen, eliminoitujen kyselyiden osalta tarkkuuden muutokset parempaan suuntaan ovat yleisiä, näin käy eliminoidussa kannassa 20 ja eliminoimattomassa kannassa 21 hakuaiheessa. Näissä kyselyissä tarkkuudet myös pysyvät useammin samoina, 8 hakuaiheessa molemmissa kannoissa, verrattuna toiseen yhdyssanamenetelmään, jossa näin käy ainoastaan yhdessä hakuaiheessa. Hakuaihekohtainen analyysi antaa viitteitä siitä, että myös CLEF2002-aineiston kohdalla yhdyssanojen osittaminen eliminointia hyödyntäen on toista yhdyssanamenetelmää parempi.

TAULUKKO 24. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn (CLEF2002, rakenteettomat kyselyt)

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	21	20
Huonontunut	20	14
Sama	1	8
Ei yhdyssanoja	7	7
Yhteensä	49	49
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	22	21
Huonontunut	19	13
Sama	1	8
Ei yhdyssanoja	7	7
Yhteensä	49	49

Myös CLEF2002-hakuaiheiden rakenteettomien kyselyiden joukosta löytyy esimerkkejä kompositionaalisista yhdyssanoista, joita ei kannata osittaa. Näin on esimerkiksi hakuaiheissa 94 (**nobelpristagare**), 97 (**folkomröstning**) ja 102 (**skidtävling**). Hakuaiheessa 94 esiintyy tosin myös muitakin yhdyssanoja kuin kompositionaalisia. Myös rakenteisten kyselyjen osalta havaittiin, ettei hakuaiheen 102 yhdyssanaa **skidtävling** kannata osittaa. CLEF2002-aineiston rakenteettomien kyselyjen joukossa on myös runsaasti kompositionaalisia yhdyssanoja, joiden osittaminen on hyödyllistä. Esimerkiksi hakuaiheissa 101 (**bronkialastma**, **luftrörssjukdom**), 111 (**datoranimering**, **filmindustri**), 137 (**skönhetstävling**) ja rakenteisten kyselyjen tavoin hakuaiheessa 139 (**fiskekvot**).

8.2.3 Ahlgren-kyselysarjoilla saadut tulokset

TAULUKKO 25. Keskimääräiset ei-interpoloidut tarkkuusarvot (%) rakenteettomien Ahlgren-kyselysarjojen osalta.

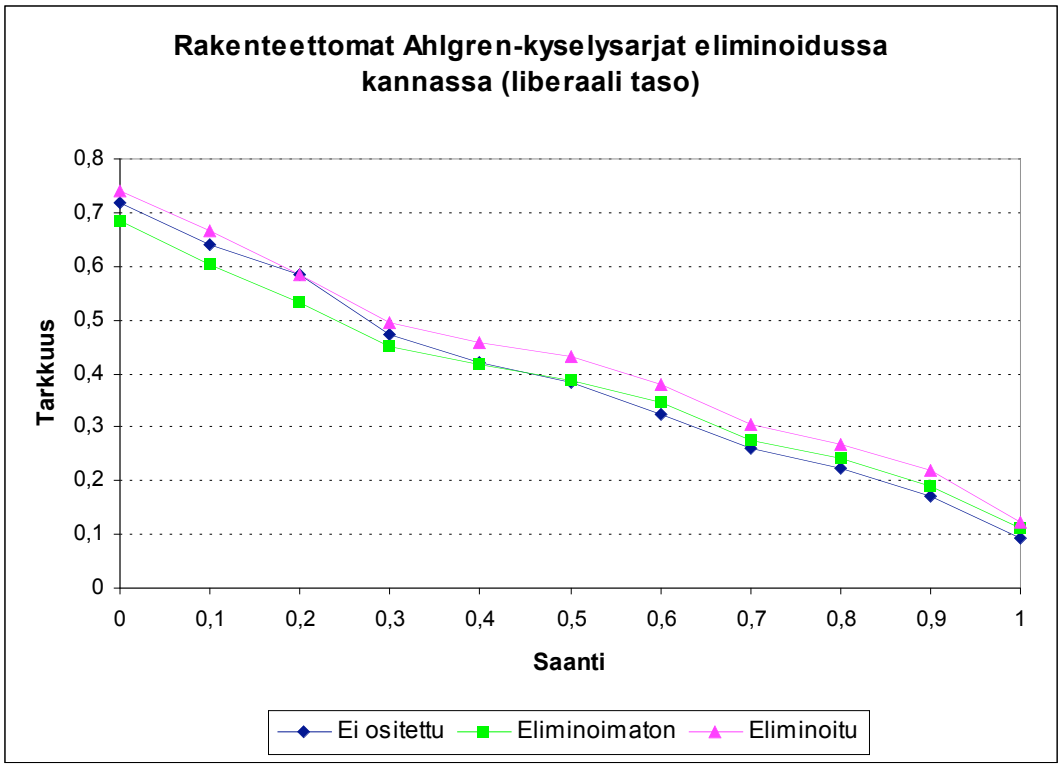
ELIMINOITU KANTA, kyselyt	Liberaali taso	Normaali taso	Tiukka taso
Ei ositettu	37,28	33,5	27,86
Ositettu, ei eliminoitu	36,93	30,66	26,31
Ei ositetun ja ositetun ero (%-yks.)	-0,35	-2,84	-1,55
Ositettu, eliminoitu	41,03	36,62	31,37
Ei ositetun ja eliminoidun ero (%-yks.)	3,75	3,12	3,51
ELIMINOIMATON KANTA, kyselyt	Liberaali taso	Normaali taso	Tiukka taso
Ei ositettu	34,4	30,74	26,31
Ositettu, ei eliminoitu	35,97	28,52	24,91
Ei ositetun ja ositetun ero (%-yks.)	1,57	-2,22	-1,4
Ositettu, eliminoitu	38,56	33,87	29,69
Ei ositetun ja eliminoidun ero (%-yks.)	4,16	3,13	3,38

Taulukossa 25 on esitetty tulokset rakenteettomilla Ahlgren-kyselysarjoilla tehdyistä eräajoista. Ahlgrenin kokoelmassa käytetään moniportaisia relevanssiarvioita, joten tuloksiakin on mahdollista tarkastella kolmella eri relevanssitasolla. Liberaalilla relevanssitasolla molemmissa kannoissa parhaiten menestyvät ositetut ja eliminoidut kyselysarjat. Eliminoidussa kannassa tarkkuuden parannus on 3,75 ja eliminoimattomassa kannassa 4,16 prosenttiyksikköä. Toinen yhdyssanamenetelmä, osittaminen eliminointia hyödyntämättä, saa eliminoidussa kannassa osittamattoman kyselysarjan kanssa samankaltaisia tarkkuusarvoja (ero vain 0,35 prosenttiyksikköä) ja eliminoimattomassa kannassa tarkkuusarvo paranee 1,57 prosenttiyksikköä.

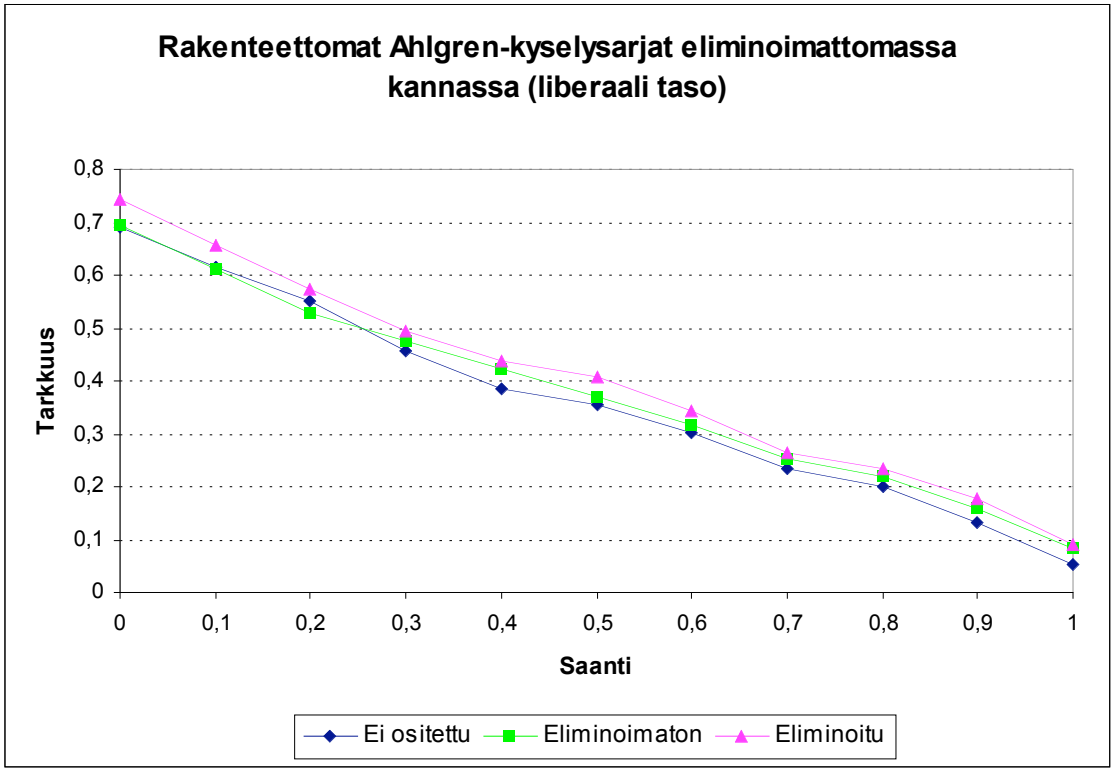
Normaalilla relevanssitasolla paras menetelmä on myös yhdyssanojen osittaminen eliminointia hyödyntäen. Toinen yhdyssanamenetelmä eli yhdyssanojen osittaminen ilman eliminointia puolestaan on molemmissa kannoissa tarkkuusarvoiltaan 2,84 ja 2,22 prosenttiyksikköä osittamatonta kyselysarjaa huonompi. Tiukalla relevanssitasolla tulokset ovat normaalin tason kaltaiset. Yhdyssanoiltaan ositetut ja eliminoimattomat kyselysarjat saavat huonoimpia tarkkuusarvoja, eliminoidut kyselysarjat parhaimpia. Tilastollisissa testeissä ei kuitenkaan havaittu merkitseviä eroja rakenteettomien kyselysarjojen tuloksissa. Kun verrataan tarkkuusarvoja rakenteisilla kyselysarjoilla saatuihin tarkkuusarvoihin, huomataan, että menetelmien paremmuusjärjestys on sa-

ma, mutta tarkkuusarvojen väliset erot ovat pienempiä rakenteisten kyselysarjojen tuloksissa. Tilastollisissa testeissä havaittiinkin merkitsevä ero ($p<0,05$) rakenteisten ja rakenteettomien eliminoitujen kyselyjen tarkkuuksissa eliminoidussa kannassa liberaalilla relevanssitasolla.

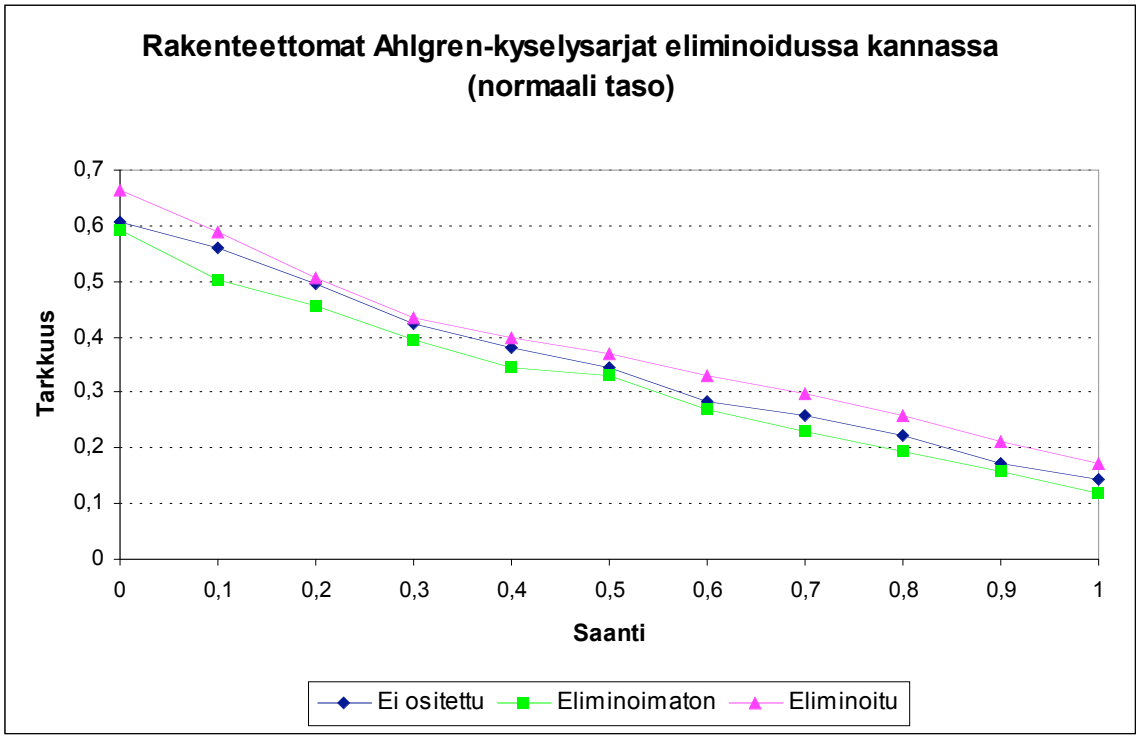
Myös kuvioista 23, 24, 25, 26, 27 ja 28 huomataan, että yhdyssanojen osittaminen eliminointia hyödyntäen on kaikilla relevanssitasoilla parhaiten menestyvä menetelmä. Yhdyssanojen osittaminen eliminointia hyödyntämättä on liberaalia tasoa lukuun ottamatta huonoiten menestyvä menetelmä.



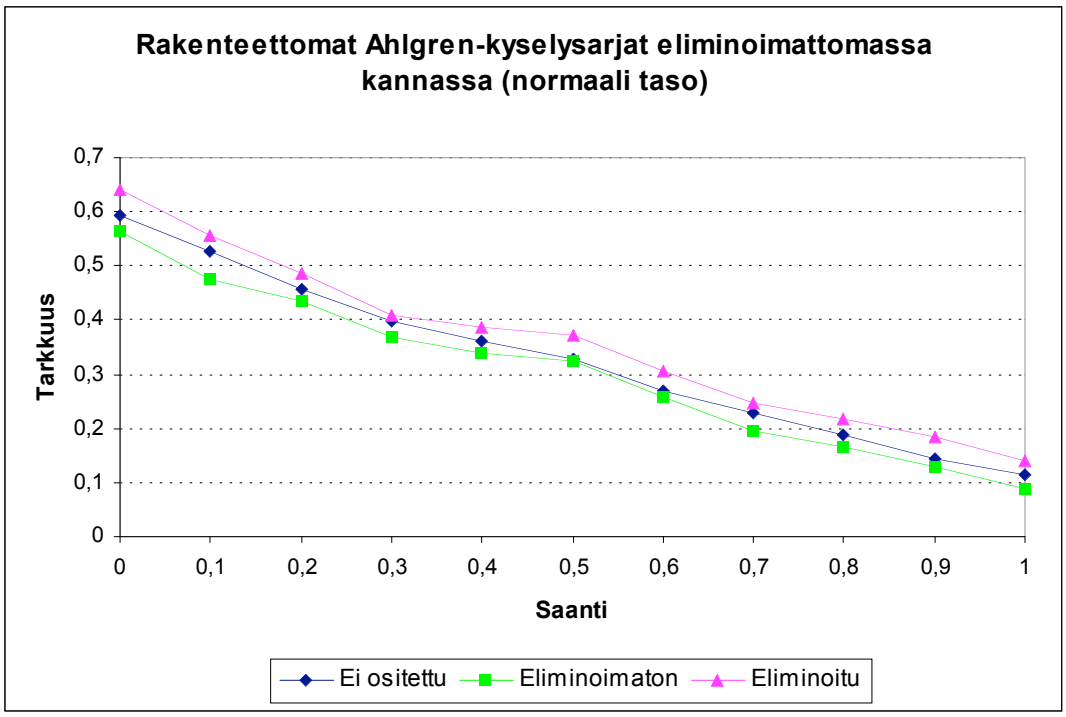
KUVIO 23. Rakenteettomilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa liberaalilla relevanssitasolla.



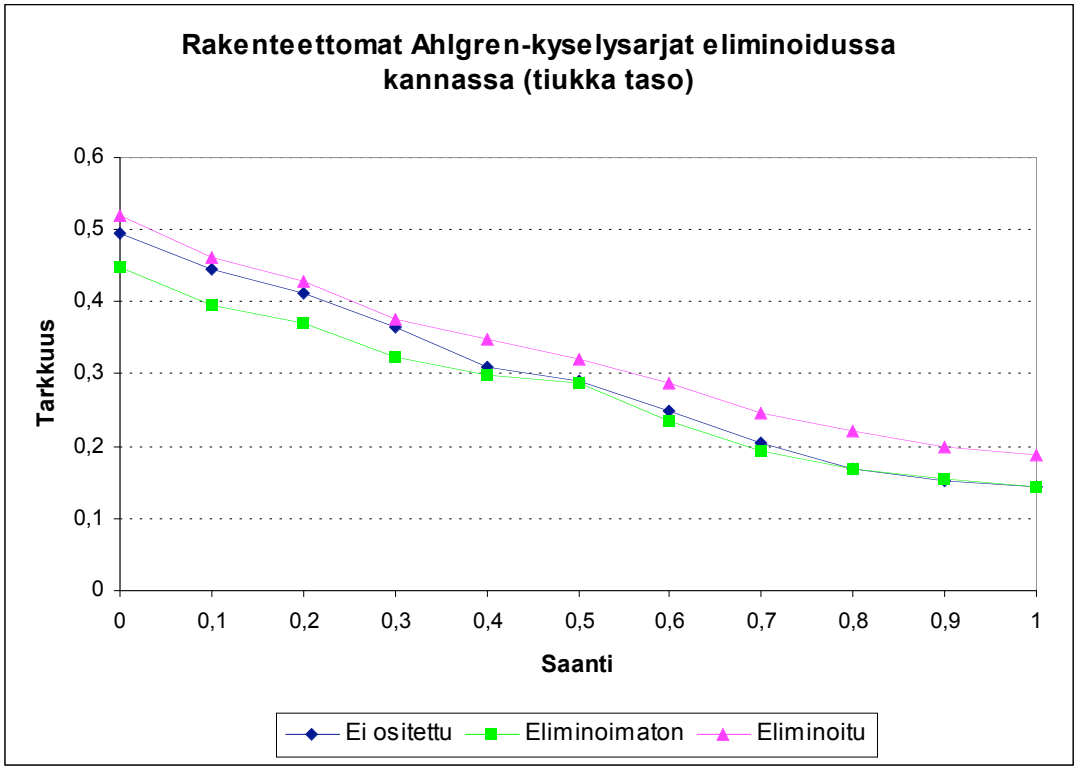
KUVIO 24. Rakenteettomilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa liberaalilla relevanssitasolla.



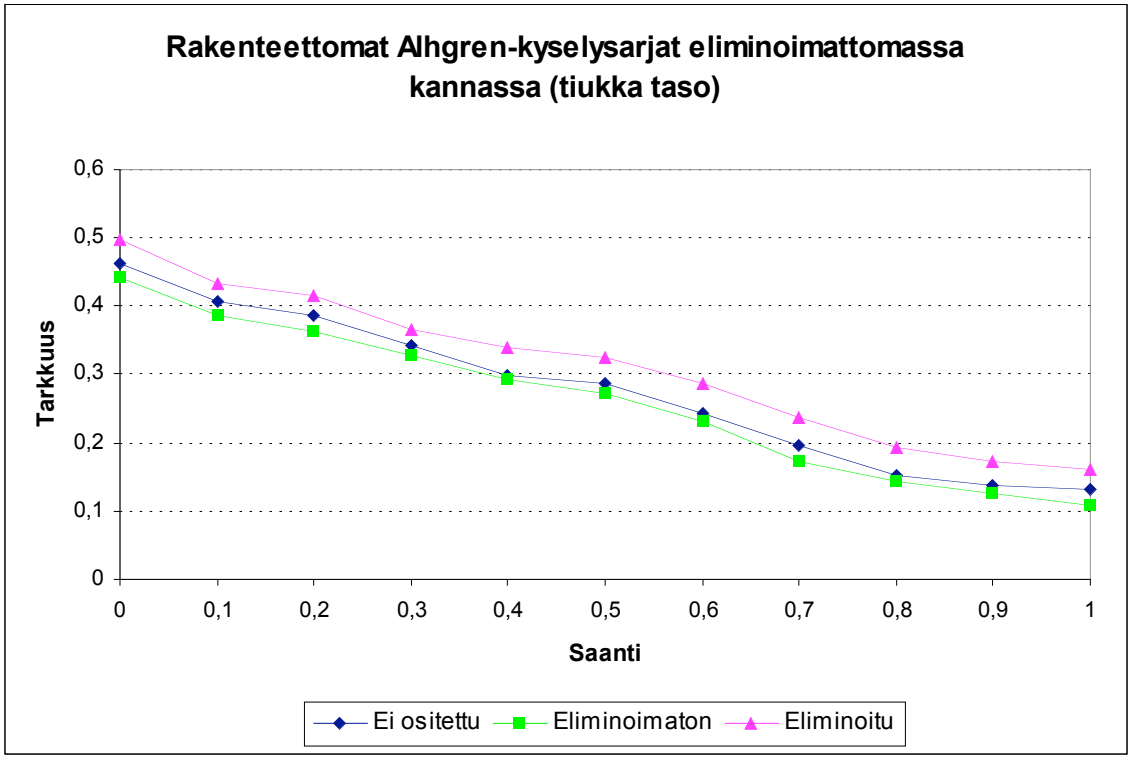
KUVIO 25. Rakenteettomilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa normaalilla relevanssitasolla.



KUVIO 26. Rakenteettomilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa normaalilla relevanssitasolla.



KUVIO 27. Rakenteettomilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoidussa kannassa tiukalla relevanssitasolla.



KUVIO 28. Rakenteettomilla Ahlgren-kyselysarjoilla saadut tarkkuudet eliminoimattomassa kannassa tiukalla relevanssitasolla.

Taulukossa 26 on esitetty hakuaihekohtaisten tarkkuuksien muutokset rakenteettomien Ahlgren-kyselyiden osalta liberaalilla relevanssitasolla. Yhteensä 51 hakuaiheesta 3 hakuaiheessa ei ole lainkaan yhdyssanoja. Tarkkuuden muutokset jakautuvat varsin tasaisesti. Ositettujen, ei eliminoitujen kyselyiden osalta tapahtuu enemmän tarkkuuden huonontumista. Molemmissa kannoissa jopa 27 hakuaiheen osalta tarkkuudet ovat huonontuneet yhdyssanojen osittamisen myötä. Ositettujen ja eliminoitujen kyselyiden tarkkuudet ovat enemmistössä parantuneet, eliminoidussa kannassa 21 ja eliminoimattomassa kannassa 22 hakuaiheessa. Näiden kyselyiden osalta tarkkuudet ovat myös useammin pysyneet samoina, molemmissa kannoissa 11 hakuaiheessa, verrattuna toiseen yhdyssanamenetelmään, jossa tarkkuudet eivät ole pysyneet samana yhdessäkään hakuaiheessa.

TAULUKKO 26. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn, rakenteettomat Ahlgren-kyselyt, liberaali taso

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	21	21
Huonontunut	27	16
Sama	0	11
Ei yhdyssanoja	3	3
Yhteensä	51	51
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	21	22
Huonontunut	27	15
Sama	0	11
Ei yhdyssanoja	3	3
Yhteensä	51	51

Taulukossa 27 on puolestaan esitetty samat tarkkuusarvojen muutokset normaalilla relevanssitasolla. Normaalilla relevanssitasolla tulokset saatiin 50 hakuaiheen kyselyille, ja 3 hakuaiheessa ei esiinny lainkaan yhdyssanoja. Muutokset ovat samankaltaisia kuin liberaalilla tasolla, eli ositettujen, ei eliminoitujen kyselyiden osalta tarkkuus huononee eliminoidussa kannassa 27 ja eliminoimattomassa kannassa 28 hakuaiheen osalta. Ositettujen ja eliminoitujen kyselyiden osalta yleisempää on tarkkuuden parantuminen, eliminoidun kannan 20 ja eliminoimattoman kannan 21 hakuaiheessa. Myös normaalilla tasolla eliminoitujen kyselyiden tarkkuudet pysyvät eliminoimattomia kyselyitä useammin samoina, molemmissa kannoissa 11 hakuaiheessa.

TAULUKKO 27. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn, rakenteettomat Ahlgren-kyselyt, normaali taso

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	20	20
Huonontunut	27	16
Sama	0	11
Ei yhdyssanoja	3	3
Yhteensä	50	50
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	19	21
Huonontunut	28	15
Sama	0	11
Ei yhdyssanoja	3	3
Yhteensä	50	50

Taulukossa 28 on puolestaan esitetty hakuaihekohtaisten tarkkuusarvojen muutokset tiukalla relevanssitasolla. Tiukalla relevanssitasolla tulokset saatiin 44 hakuaiheen kyselyistä, ja kolmes-

sa hakuaiheessa ei esiinny lainkaan yhdyssanoja. Tälläkin tasolla yhdyssanojen osittaminen ilman eliminointia huonontaa tarkkuutta selvästi, eliminoidussa kannassa 24 ja eliminoimattomassa kannassa 22 hakuaiheessa. Osittaminen ja eliminoiminen puolestaan pääasiassa parantaa tarkkuutta, eliminoidun kannan 16 ja eliminoimattoman kannan 18 hakuaiheessa. Myös tiukalla tasolla eliminoitujen kyselyiden tarkkuus pysyy eliminoimattomia useammin samana, eliminoidun kannan 14 ja eliminoimattoman kannan 12 hakuaiheessa. Myös Ahlgren-kyselysarjojen osalta hakuaihekohtainen vertailu antaa viitteitä siitä, että yhdyssanojen osittaminen eliminointia hyödyntäen on toista yhdyssanamenetelmää parempi menetelmä.

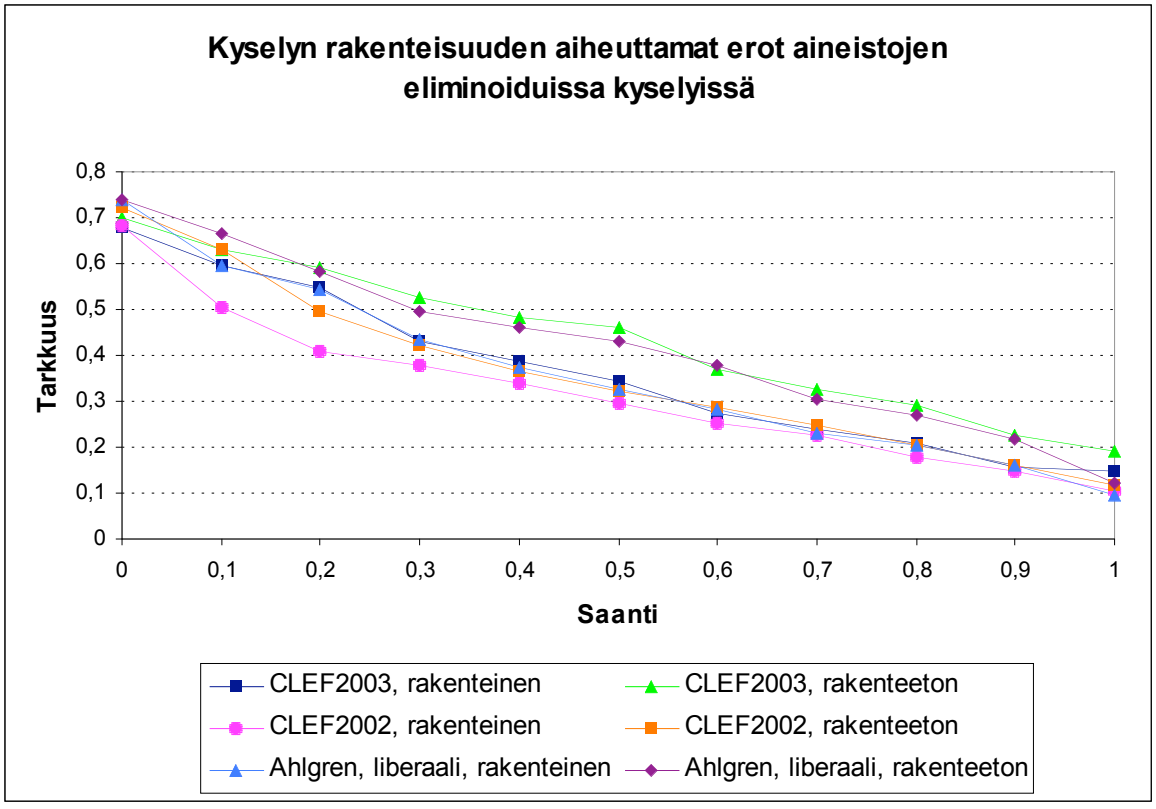
TAULUKKO 28. Hakuaihekohtaisten tarkkuuksien muutos verrattuna ei ositettuun kyselyyn, rakenteettomat Ahlgren-kyselyt, tiukka taso

ELIMINOITU KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	15	16
Huonontunut	24	11
Sama	2	14
Ei yhdyssanoja	3	3
Yhteensä	44	44
ELIMINOIMATON KANTA		
	Ositettu, ei eliminoitu	Ositettu, eliminoitu
Parantunut	18	18
Huonontunut	22	11
Sama	1	12
Ei yhdyssanoja	3	3
Yhteensä	44	44

Kun tutkitaan tarkemmin kyselyjen sisältämiä yhdyssanoja, löydetään paljon esimerkkejä hakuaiheista, joissa eliminoimaton kysely suoriutuu huonosti. Esimerkiksi hakuaiheessa 20 SWETWOL:n virhetulkintaiset yhdyssanajaot sanoissa **nackdel** (**nackedelare**) ja **gemensam** (**gemensimma**, **gem**, **simma**) voivat vaikuttaa eliminoimattoman kyselyn huonompaan menestykseen. Tämän lisäksi esiintyy esimerkkejä hakuaiheista, joissa ei ole yhdyssanoja ja jotka ovat samankaltaisia, mutta jotka saavat silti erilaisia tarkkuusarvoja. Näin oli myös rakenteisten kyselyjen samoissa hakuaiheissa. Rakenteettomista Ahlgren-kyselyistä löytyy myös paljon esimerkkejä kompositionaalisten yhdyssanojen osittamisen hyödyllisyydestä. Esimerkiksi tämä on nähtävissä hakuaiheissa 22 (**flygplansolycka**, **landningsbana**), 31 (**konsumentskydd**) ja 80 (**hungerstrejksaktion**). Rakenteisiin kyselyihin verrattuna suurimpien tarkkuusmuutosten listalla on joitakin samoja hakuaiheita, joista suurimmassa osassa tulkinnat ovat samanlaiset. Esimerkiksi hakuaiheen 61 osalta tulkinnat ovat erilaisia. Rakenteisessa kyselyssä yhdyssanan **oljeledning** osittaminen huonontaa tarkkuutta, kun taas rakenteettomassa kyselyssä tarkkuus paranee.

8.3 Kyselyn rakenteisuuden vaikutus hakutuloksiin

Edellä esitellyt tulokset antavat viitteitä siitä, että kyselyn rakenteisuudella on suuri vaikutus hakutuloksiin tässä tutkimuksessa. Rakenteettomissa kyselysarjoissa eri menetelmien saavuttamat keskimääräiset tarkkuusarvot ovat korkeampia kuin rakenteisilla kyselysarjoilla saadut tarkkuudet. Tilastollisissa testeissä erityisesti eliminoiduissa kyselysarjoissa havaittiin tilastollisesti merkitseviä eroja rakenteisten ja rakenteettomien kyselyjen tarkkuuksien välillä. CLEF2003-aineistossa tilastollisesti merkitsevä ero ($p<0,05$) havaittiin rakenteisten ja rakenteettomien eliminoitujen kyselyiden tarkkuuksien välillä eliminoidussa kannassa. Myös Ahlgren-kyselysarjoissa tilastollisesti merkitsevä ero havaittiin rakenteisten ja rakenteettomien eliminoitujen kyselyjen tarkkuuksien välillä eliminoidussa kannassa liberaalilla relevanssitasolla. CLEF2002-aineisto osoittaa jälleen poikkeavan luonteensa. CLEF2002-kyselysarjoissa tilastollisesti merkitsevä ero havaittiin rakenteisten ja rakenteettomien eliminoitujen kyselyjen tarkkuuksien välillä eliminoimattomassa kannassa.



KUVIO 29. Kyselyn rakenteisuuden aiheuttamat erot eri aineistojen eliminoiduissa kyselyissä.

Kuvio 29 esittää eri aineistoissa havaitut tilastollisesti merkitsevät erot. CLEF2002 poikkeaa kahdesta muusta aineistosta myös pienempien tarkkuusarvojen perusteella. CLEF2002-aineiston rakenteettomat kyselysarjat saavuttavat samanlaisia tarkkuuksia kuin kahden muun aineiston rakenteiset versiot. Parhaiten menestyvät CLEF2003- ja Ahlgren-aineistojen rakenteettomat eliminoidut kyselysarjat eliminoidussa kannassa.

9 TULOKSET

Tutkielman tutkimuskysymyksistä ensimmäinen ja toinen pyrkivät selvittämään yhdyssanojen määrää ja yhdyssanatyyppejä ruotsinkielisissä hakuaiheissa ja dokumenteissa. Tarkoituksena on samalla kartoittaa sitä, kuinka laajana ilmiönä yhdyssanat näyttäytyvät ruotsin kielessä. Toiseksi kiinnostuksen kohteena on se, ovatko substantiiviset yhdyssanat määrällisesti laajamittaisin yhdyssanatyyppi, vai onko olemassa muita sanaluokkia, jotka tulisi ottaa huomioon tiedonhaussa. Kolmanneksi tarkoituksena on selvittää kompositionaalisten ja ei-kompositionaalisten yhdyssanojen suhdetta aineistoissa. Tausta-ajatuksena toimii Pirkolan (2001) ajatus siitä, että mikäli aineistoissa esiintyy enemmistönä kompositionaalisia yhdyssanoja, yhdyssanojen morfologisesta käsittelystä voisi olla hyötyä tiedonhaussa. Ei-kompositionaalisten yhdyssanojen osiin jakaminen ei puolestaan ole tiedonhaun näkökulmasta tarkoituksenmukaista.

Tutkimus paljastaa yhdyssanojen monitulkintaisen luonteen. Lähtiessäni selvittämään yhdyssanojen määrää hakuaiheissa, huomasin, ettei yhdyssanasta ole olemassa johdonmukaista määritelmää. SAOL-sanakirjan yhdyssanatulkinnat ovat liberaaleja ja yhdyssanoiksi lasketaan sanoja monin eri perustein. Toinen apuväline SWETWOL osoittautui myös ongelmalliseksi virhetulkintojen suuren määrän vuoksi. Toiseksi yhdyssanojen analysoiminen osoitti, että yhdyssanojen määrä ylittää kaikissa aineistoissa sekä hakuaiheissa että dokumenteissa Hedlundin (2002) arvioiman määrän. CLEF2003- ja Ahlgren-hakuaiheissa yhdyssanojen määrä oli SAOL-yhdyssanojen osalta noin 17 prosenttia ja SWETWOL-yhdyssanojen osalta 15−16 prosenttia. CLEF2002-hakuaiheet poikkesivat kahdesta muusta aineistosta pienemmän yhdyssanamääränsä vuoksi. Myös dokumenttien osalta yhdyssanamäärä oli CLEF-dokumenttien osalta 13,6 ja Ahlgren-dokumenttien osalta 12,4 prosenttia, joten niidenkin perusteella katsottuna yhdyssanat ovat tärkeä kielenilmiö, joka tulee huomioida myös tiedonhaussa. Edellä esitetyt luvut kuvaavat Pir-

kolan (2001: 337) esittelemää yhdyssanaindeksiä (CIF-indeksiä) ruotsin kielen eri aineistojen osalta.

Sanaluokkajaon osalta kaikissa aineistoissa esiintyi eniten substantiivisia yhdyssanoja. Toiseksi eniten esiintyy yhdyssanaverbejä. Ne eivät kuitenkaan ole tiedonhaun näkökulmasta suuri ongelma, koska niitä esiintyy substantiiveihin verrattuna vähän ja koska SWETWOL tunnistaa niistä vain pienen osan. Yhdyssanojen kompositionaalisuusaste oli hakuaiheissa 60-40. Myös dokumenteissa kompositionaalisia yhdyssanoja havaittiin 60 prosentin verran, kun taas loput 40 prosenttia jakautui SWETWOL:n virhetulkintojen ja ei-kompositionaalisten yhdyssanojen kesken. Mikäli oltaisiin tarkasteltu vain substantiivisia yhdyssanoja, kompositionaalisuuden aste olisi voinut olla vieläkin suurempi, koska ei-kompositionaalisista yhdyssanoista varsin suuri osa oli yhdyssanaverbejä. Myös erisnimiyhdyssanojen määrä oli dokumenteissa yllättävän suuri, mihin ovat kiinnittäneet huomiota myös Gunnarsson ja Petersson (2005). Sekä erisnimiyhdyssanat että yhdyssanaverbit ovat luonteeltaan sellaisia, ettei niitä kannata jakaa osiin tiedonhaussa. Kompositionaalisten yhdyssanojen enemmistöosuus antoi kuitenkin viitteitä siitä, että yhdyssanojen käsittely voisi olla ruotsin kielessä tarkoituksenmukaista.

Tutkimuksen tutkimuskysymyksistä kolmannella pyritään puolestaan hakemaan vastausta juuri siihen, kannattaako yhdyssanoja käsitellä morfologisesti ruotsinkielisissä kyselyissä. Aiemman tutkimuksen valossa oli oletettavaa, että yhdyssanojen osittaminen olisi ruotsin kielen kannalta hyödyllistä. Tässä tutkimuksessa morfologisella käsittelyllä tarkoitetaan kahta eri menetelmää. Ensimmäinen menetelmistä on yhdyssanojen osittaminen, jota verrattiin perusmuotoiseen kyselysarjaan, jonka yhdyssanoille ei ole tehty mitään. Toinen menetelmä hyödyntää yhdyssanojen osittamisen lisäksi eliminointiperiaatetta, eli yhdyssanojen osittamisen tuottamia yhdysosia eliminoidaan. Tämä osa tutkimuksesta on tiedonhaun laboratoriotutkimus. Tutkimuksessa on käytössä kaksi erilaista hakemistoa, eliminoitu ja eliminoimaton kanta, joihin tiedonhaut kohdistettiin kolmella erilaisella kyselysarjalla. Tämän lisäksi kyselysarjoista oli käytössä sekä rakenteettomat että rakenteiset versiot.

Rakenteisten kyselysarjojen osalta kolme eri aineistoa antoivat ristiriitaisia tuloksia sen suhteen, kannattaako yhdyssanoja käsitellä ruotsinkielisissä kyselyissä. Yhdyssanojen käsittelylle myönteisin aineisto oli CLEF2003, jossa yhdyssanojen osittaminen ilman eliminointia paransi tarkkuusarvoja sekä eliminoidussa että eliminoimattomassa kannassa. Myös yhdyssanojen osittaminen eliminointia hyödyntäen paransi tarkkuutta, joskin erot menetelmien välillä olivat varsin

pieniä. Toinen CLEF-aineisto, CLEF2002, suhtautui täysin päinvastaisella tavalla yhdyssanojen käsittelyyn. Molemmat yhdyssanojen käsittelymenetelmät huononsivat tarkkuuksia sekä eliminoidussa että eliminoimattomassa kannassa. Ahlgren-kyselysarjojen osalta tulokset olivat myös ristiriitaisia. Yhdyssanojen osittaminen ilman eliminointia huononsi tarkkuuksia kaikilla relevanssitasoilla lukuun ottamatta eliminoimattoman kannan eliminoimatonta kyselysarjaa liberaalilla relevanssitasolla. Yhdyssanojen osittaminen ja eliminoiminen puolestaan paransi tarkkuusarvoja, mutta erot yhdyssanoiltaan osittamattomaan kyselysarjaan olivat marginaaliset, varsinkin tiukalla relevanssitasolla erot jäivät hyvin pieniksi. Menetelmien välillä ei havaittu yhdessäkään aineistossa tilastollisesti merkitseviä eroja.

Ruotsin kielessä on siis kompositionaalisia yhdyssanoja enemmän, mutta rakenteisilla kyselysarjoilla saadut tulokset viittaavat siihen, että yhdyssanojen kompositionaalisuus ei ole tae sille, että yhdyssanojen käsittely on kielessä tuloksellista. Myös rakenteisten kyselyjesarjojen hakuaihekohtaisia tarkkuuksia tutkimalla saatiin viitteitä siitä, että hakuaiheiden välillä on suuria eroja sen suhteen, kannattaako yhdyssanoja käsitellä vai ei. Yhdyssanojen luonne voi kompositionaalisuudesta huolimatta vaihdella paljon, ja myös kompositionaalisen yhdyssanan osittaminen voi johtaa huonoihin lopputuloksiin.

Tämä näyttäisi olevan yhteydessä Ahlgrenin (2004: 131–133) arvioon siitä, että yhdyssanojen osittamisen tehokkuus hakumenetelmänä on hakuaihekohtaista ja riippuvaista hakuaiheen aiheesta. Rakenteisten kyselyjen hakuaihekohtaisessa vertailussa löytyi esimerkkejä siitä, että kompositionaalisen yhdyssanan osittaminen parantaa hakutulosta (esimerkiksi CLEF2002-hakuaiheen 139 **fiskekvot**). Monissa tapauksissa yhdyssanojen osittaminen ei ollut hyödyllistä juuri yhdyssanojen luonteen vuoksi. Yhdyssanojen osittamisen myötä hakuun tuli mukaan liian yleisiä hakuavaimia (esimerkiksi CLEF2003-hakuaiheen 142 **riksdagshuset**). Joissakin hakuaiheissa SWETWOL:n tekemä virhetulkinta huononsi tarkkuuksia (esimerkiksi CLEF2003-hakuaiheen 196 **bankerna**). Esimerkiksi oli havaittavissa tapauksia, joissa SWETWOL ensin palautti sanan vääränlaiseen perusmuotoon, mikä myös johti siihen, että yhdyssanojen osittaminen epäonnistui. Rakenteisten kyselysarjojen tuloksia tarkasteltaessa on kuitenkin syytä pitää mielessä se, että myös #uw20-läheisyysoperaattorin käytöllä voi olla vaikutusta saatuihin tuloksiin, koska operaattori rajaa kyselyitä paljon ja edellyttää, että molemmat yhdysosat esiintyvä 20 sanan etäisyydellä toisistaan. Tämän vuoksi tutkimuksessa tehtiin myös eräajot rakenteettomilla kyselyillä, joissa kyselyt koostuvat sanalistoista, jotka yhdistää toisiinsa ainoastaan #combine-operaattori.

Rakenteettomilla kyselysarjoilla saadut tulokset ovat eriluonteisia kuin rakenteisilla kyselysarjoilla saadut tulokset. Rakenteisten ja rakenteettomien eliminoitujen kyselyiden välillä havaittiin myös tilastollisesti merkitseviä eroja. Rakenteettomien kyselysarjojen tulosten perusteella yhdyssanojen morfologisesta käsittelystä voisi olla hyötyä ruotsinkielisessä tiedonhaussa. Kaikkien aineistojen osalta parhaiten menestyvä menetelmä oli molemmissa kannoissa yhdyssanojen osittaminen eliminointiperiaatetta hyödyntäen. Myös rakenteettomien kyselyiden osalta yhdyssanojen käsittelylle myönteisin aineisto oli CLEF2003. Rakenteisista kyselysarjoista poiketen myös CLEF2002-aineistossa havaittiin tarkkuusarvojen parantumista yhdyssanojen käsittelyn myötä. Ahlgren-kyselysarjojen osalta tulokset olivat rakenteisten kyselysarjojen tavoin hieman ristiriitaisia. Yhdyssanojen osittaminen eliminointia hyödyntäen oli tämänkin aineiston osalta tehokkain menetelmä, mutta ositetut ja eliminoimattomat kyselysarjat saivat huonoimpia tarkkuusarvoja. Rakenteettomien kyselysarjojen tuloksissa menetelmien välillä ei havaittu tilastollisesti merkitseviä eroja yhdessäkään aineistossa.

Rakenteettomien kyselysarjojen hakuaihekohtaisten tarkkuuksien tarkastelu osoitti myös suurta vaihtelua eri hakuaiheiden välillä, joskin varsinkin CLEF2003-aineiston osalta enemmistö tarkkuuden muutoksista tapahtui positiiviseen suuntaan. Muissakin aineistoissa varsinkin yhdyssanojen osittaminen eliminointia hyödyntäen sai aikaan enemmän positiivisia kuin negatiivisia tarkkuuden muutoksia, joskin erot olivat CLEF2003-aineistoon verrattuna pienempiä. Kun siis tarkastellaan asiaa kyselyn rakenteisuuden näkökulmasta, rakenteettomat kyselysarjat menestyivät kaikin puolin rakenteisia kyselysarjoja paremmin. Myös rakenteettomista kyselyistä löytyi esimerkkejä kompositionaalisista yhdyssanoista, joiden osiin jakaminen ei paranna tarkkuuksia (esimerkiksi CLEF2003-hakuaiheen 163 **regelverk** ja **lagstiftning**). Kyselyissä esiintyi myös paljon SWETWOL:n virhetulkintoja, jotka huononsivat tarkkuusarvoja erityisesti yhdyssanoiltaan eliminoimattomissa kyselyissä (esimerkiksi CLEF2003-hakuaiheen 180 **konkurs**). Pääasiassa vaikuttaa kuitenkin siltä, että kompositionaalisten yhdyssanojen osiin jakaminen on tarkoituksenmukaista. Tällaisia hakuaiheita esiintyi paljon (esimerkiksi CLEF2002-hakuaiheen 111 **datoranimering** ja **filmindustri**).

10 PÄÄTELMÄT

Tutkimuksessa on selvitetty yhdyssanojen roolia tiedonhaussa ruotsin kielessä. Tutkimus osoitti, että yhdyssanat ovat tärkeä kielenilmiö, jota ei ole syytä unohtaa myöskään kyselyjä muodostettaessa. Yhdyssanat ovat kuitenkin myös monitulkintaisia, mikä osaltaan voi selittää sitä, että kyselynmuodostusmenetelmien välillä saatiin ristiriitaisia tuloksia eri aineistoissa rakenteisten ja rakenteettomien kyselysarjojen osalta. Rakenteisten kyselysarjojen huonoon menestykseen voi tosin vaikuttaa myös läheisyysoperaattorien käyttö. Kun verrataan toisiinsa rakenteisilla ja rakenteettomilla kyselysarjoilla saatuja tuloksia, ei varmuudella tiedetä, mikä on läheisyysoperaattorien vaikutus rakenteisten kyselysarjojen huonompiin tarkkuusarvoihin. On mahdollista, että #uw20-läheisyysoperaattori on ehdoiltaan liian tiukka ja rajaa hakua liikaa. Tämän vuoksi tarkkuudetkin kärsivät. Ei myöskään ole varmuutta siitä, mikä on läheisyysoperaattoria käytettäessä optimaalinen sanojen välinen etäisyys. On mahdollista, että tulokset olisivat parempia, jos olisi käytetty 20 sanaa pienempää etäisyyttä. Rakenteettomien kyselysarjojen osalta voitaisiin kuitenkin tehdä sellainen johtopäätös, että yhdyssanojen morfologinen käsittely kannattaa ruotsin kielessä, joskaan menetelmien väliset erot eivät olleet tilastollisesti merkitseviä. Myös aikaisemmat tutkimukset (Ahlgren & Kekäläinen 2006; Kettunen 2007; Hollink 2004) ovat havainneet yhdyssanojen käsittelyn hyödylliseksi ruotsin kielessä. Myös saksan kielessä (Braschler & Ripplinger 2004) yhdyssanojen käsittely on havaittu hyödylliseksi menetelmäksi.

Tutkimuksessa on kartoitettu yhdyssanojen kompositionaalisuuden astetta ruotsin kielessä. Ruotsin kielen osalta kompositionaalisten yhdyssanojen osuus on sekä hakuaiheissa että dokumenteissa 60 prosenttia. Kuitenkaan ei ole täysin selvää, onko ruotsin kielen yhdyssanojen käsittelystä hyötyä tiedonhaussa, koska eri aineistot antavat ristiriitaisia tuloksia. Rakenteettomien kyselysarjojen tulosten perusteella voitaisiin vetää Pirkolan (2001) idean mukainen johtopäätös. Kun kompositionaalisia yhdyssanoja on enemmän, yhdyssanojenkin käsittely on kannattavaa. Jatkossa olisikin mielenkiintoista tehdä kielten välistä vertailua. Mikä on esimerkiksi suomen tai saksan kielten yhdyssanojen kompositionaalisuuden aste? Kielet muistuttavat toisiaan ainakin yhdyssanojen runsaan esiintymisen perusteella. Voidaanko löytää myös muunlaisia yhtymäkohtia yhdyssanatyyppien ja niiden morfologisen käsittelyn suhteen? Pirkola esittää erilaisten indeksien laskemista eri kielille niiden morfologisten ominaisuuksien osalta. Indeksien avulla kielten välinen vertailu olisi helpompaa ja yhden kielen tutkimustuloksia voitaisiin soveltaa paremmin toiseen kieleen, joka on indeksiltään samanlainen.

Koska hakuaihekohtaisessa vertailussa havaittiin molempien kyselysarjojen osalta paljon vaihtelua hakuaiheen luonteen ja yhdyssanojen luonteen kohdalla, myös yhdyssanojen valikoiva osittaminen voisi olla ruotsin kielessä kannattava menetelmä. Näin voitaisiin jakaa osiin vain sellaiset yhdyssanat, joiden kohdalla osittaminen todella on hyödyllistä. Tätäkin menetelmää voisi jatkossa testata ruotsinkielisessä tiedonhaussa.

Tässä tutkielmassa on keskitytty ruotsin kielen ominaispiirteistä yhteen eli yhdyssanojen laajamittaiseen esiintymiseen kielessä. Ruotsin kieleen liittyy kuitenkin myös toinen kielellinen ilmiö, joka vaatisi huomioon ottamista. Jatkossa voisikin olla hyödyllistä tutkia ruotsin kielen homografiaongelmaa. 65 prosenttia ruotsinkielisen tekstin sananmuodoista on homografeja. Myös tämän tutkimuksen aineistoissa havaittiin esimerkkejä homografisista sanoista, jotka aiheuttavat morfologiselle analyysiohjelmalle ongelmia. Olisikin mielenkiintoista selvittää tarkemmin, millaisena tämä ilmiö näyttäytyy tiedonhaussa ja millaisia vaikutuksia ilmiöllä on esimerkiksi morfologisen analyysiohjelman SWETWOL:n toimintaan.

Aineiston analyysi antoi myös lukuisia esimerkkejä siitä, että SWETWOL:n tekemä yhdyssanojen analyysi ei vielä toimi täydellisesti. Tutkielmassa voidaan kuitenkin yhtyä Ahlgrenin ja Kekäläisen (2006) johtopäätökseen siitä, että sanojen morfologinen variaatio on otettava ruotsin kielessä huomioon jollakin tavalla. Tiedonhakijan toiminnan helpottamisen näkökulmasta olisikin kannattavinta, että variaatioon puututaan dokumenttien tallennusvaiheessa. Jollei näin toimita, tiedonhakijan on otettava kaikki tämä variaatio huomioon muotoillessaan kyselyitä.

LÄHTEET

Ahlgren, P. 2004. The Effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database. Valfrid 28. Väitöskirja. Borås: University College of Borås/Göteborg University.

Ahlgren, P. & Kekäläinen, J. 2006. Swedish full text retrieval: Effectiveness of different combinations of indexing strategies with query terms. Information Retrieval 9 (6), 681–697. Saatavilla SpringerLink-verkkolehtipalvelusta <http://springerlink.metapress.com> (käytetty 22.3.2007).

Airio, E. 2006. Word normalization and decompounding in mono- and bilingual IR. Information Retrieval 9 (3), 249–271. Saatavilla SpringerLink-verkkolehtipalvelusta <http://springerlink.metapress.com> (käytetty 22.3.2007).

Alaterä, A. & Halttunen, K. 2002. Tiedonhaun perusteet – osa lukutaitoa. Helsinki: BTJ Kirjastopalvelu Oy.

Alkula, R. & Honkela, T. 1992. Tekstin tallennus- ja hakumenetelmien kehittäminen suomen kielen tulkintaohjelmien avulla. FULLTEXT-projektin loppuraportti. VTT julkaisuja 765. Espoo: Valtion teknillinen tutkimuskeskus.

Alkula, R. 2000. Merkkijonoista suomen kielen sanoiksi. Suomen kielen morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja liittämisen vaikutukset tekstin tallennukseen ja hakuun. Acta Universitatis Tamperensis 763. Väitöskirja. Tampere: Tampereen yliopisto.

Allén, S. et al. 1980. Nusvensk frekvensordbok baserad på tidningstext. Osa 4: Ordled, betydelser. Stockholm: Almqvist & Wiksell.

Baeza-Yates, R. & Ribeiro-Neto, B. 1999. Modern information retrieval. New York: ACM Press.

Blåberg, O. 1988. A study of Swedish compounds. Umeå: University of Umeå.

Braschler, M. & Ripplinger, B. 2004. How effective is stemming and decompounding for German text retrieval? Information Retrieval 7 (3–4), 291–316. Saatavilla SpringerLink-verkkolehtipalvelusta <http://springerlink.metapress.com> (käytetty 22.3.2007).

Carlson, L. & Honkela, T. 1993. Luonnollisen kielen käsittely. Teoksessa Eero Hyvönen, Ilkka Karanta ja Markku Syrjänen. (toim.) Tekoälyn ensyklopedia. Helsinki: Gaudeamus. S. 233–243.

Carlberger, J. et al. 2001. Improving precision in information retrieval for Swedish using stemming. Teoksessa Proceedings of NODALIDA '01 – 13th Nordic Conference on Computational Linguistics. Saatavilla verkosta <ftp://ftp.nada.kth.se/pub/documents/IPLab/TechReports/IPLab-194.pdf> (käytetty 11.10.2007).

Cross Language Evaluation Forum. 2008. <http://www.clef-campaign.org/> [www] (käytetty 30.3.2008).

Gunnarsson, D. & Petersson, C. 2005. Queryexpansion med böjningsvarianter och uppbrytning av sammansättningar. Magisteruppsats. Borås: Högskola i Borås. Saatavilla verkosta <http://hdl.handle.net/2320/1358> (käytetty 11.10.2007).

Hedlund, T., Pirkola, A. & Järvelin, K. 2001. Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. Information Processing & Management 37 (1), 147–161. Saatavilla ScienceDirect-tietokannasta <http://www.sciencedirect.com/> (käytetty 22.3.2007).

Hedlund, T. 2002. Compounds in dictionary-based cross-language information retrieval. Information Research 7 (2) January 2002. Saatavilla verkosta <http://informationr.net/ir/7-2/paper128.html> [ei sivunumerointia] (käytetty 22.3.2007).

Hedlund, T. 2003. Dictionary-based cross-language information retrieval. Principles, system design and evaluation. Acta Universitatis Tamperensis 962. Väitöskirja. Tampere: University of Tampere.

Hellberg, Staffan. 1978. The morphology of present-day Swedish. Word-inflection, word-formation, basic dictionary. Stockholm: Almquist & Wiksell International.

Hollink, V. et. al. 2004. Monolingual document retrieval for European languages. Information Retrieval 7 (1–2), 33–52. Saatavilla SpringerLink-verkkolehtipalvelusta <http://springerlink.metapress.com> (käytetty 13.1.2007).

Holopainen, M. & Pulkkinen, P. 1995. Tilastolliset menetelmät. Perusteet. Porvoo: Weilin + Göös.

Hultman, T. G. 2003. Svenska Akademiens språklära. Stockholm: Svenska Akademien.

Häkkinen, K. 2001. Kielitieteen perusteet. Helsinki: Suomalaisen Kirjallisuuden Seura.

Ingwersen, P. 1992. Information retrieval interaction. London: Taylor Graham.

Ingwersen, P. & Järvelin, K. 2005. The Turn. Integration of information seeking and retrieval in context. Dordrecht: Springer.

Järvelin, K. 1995. Tekstitiedonhaku tietokannoista. Johdatus periaatteisiin ja menetelmiin. Espoo: Suomen ATK-kustannus.

Järvelin, K. & Sormunen, E. 1999. Dokumentit kateissa: tiedon tallennus ja haku avuksi. Teoksessa Ilkka Mäkinen (toim.) Tiedon tie: johdatus informaatiotutkimukseen. Helsinki: BTJ Kirjastopalvelu. S. 110–143.

Karlsson, F. 1992. SWETWOL: A Comprehensive morphological analyser for Swedish. Nordic Journal of Linguistics 15 (1), 1–45.

Karlsson, F. 2006. Yleinen kielitiede. Helsinki: Yliopistopaino.

Kettunen, K. 2007. Reductive and generative approaches to morphological variation of keywords in monolingual information retrieval. Acta Universitatis Tamperensis 1261. Väitöskirja. Tampere: University of Tampere. Saatavilla verkossa osoitteessa <http://acta.uta.fi/pdf/978-951-44-7088-2.pdf > (käytetty 8.3.2008).

Koskenniemi, K. 1983. Two-level morphology: A General computational model for word-form recognition and production. Publications no. 11. Väitöskirja. Helsinki: University of Helsinki.

Leppänen, E. 1995. Homografien disambiguoinnin vaikutukset ja toteuttaminen teksti-tiedonhaussa. Pro gradu -tutkielma. Tampere: Tampereen yliopisto.

Liljestrand, B. 1993. Så bildas orden. Handbok i ordbildning. Lund: Studentlitteratur.

Malmgren, S. 1994. Svensk lexikologi. Ord, ordbildning, ordböcker och orddatabaser. Lund: Studentlitteratur.

MOT Atk-sanakirja 1.0. 2007. Kielikone. Saatavilla lisenssiä vastaan verkosta <http://mot.kielikone.fi> [www] (käytetty 22.2.2008).

Pickard, A. J. 2007. Research methods in information. London: Facet Publishing.

Pirkola, A. 1999. Studies on linguistic problems and methods in text retrieval. The Effects of anaphor and ellipsis resolution in proximity searching, and translation and query structuring methods in cross-language retrieval. Acta Universitatis Tamperensis 672. Väitöskirja. Tampere: University of Tampere.

Pirkola, A. 2001. Morphological typology of languages for IR. Journal of Documentation, 57 (3), 330–348. Saatavilla Emerald-tietokannasta <http://www.emeraldinsight.com> (käytetty 22.3.2007).

Robertson, S. E. 1981. The methodology of information retrieval experiment. Teoksessa K. Sparck Jones et al. (toim.) 1981. Information retrieval experiment. London: Butterworths.

Sanderson, M. 1996. Word sense disambiguation and information retrieval. Väitöskirja. University of Glasgow. Department of computer science. Saatavilla verkossa osoitteessa: <http://dis.shef.ac.uk/mark/cv/publications/papers/my_papers/PhD_Thesis.pdf> (käytetty 24.2.2008).

Saracevic, T. 1996. Relevance reconsidered. Teoksessa: Information science: Integration in perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2). Copenhagen (Denmark), 14-17 Oct.1996. S. 201-218. Saatavissa verkosta osoitteessa: <www.scils.rutgers.edu/~tefko/CoLIS2_1996.doc> (käytetty 21.3.2008).

Smucker, M. D., Allan, J. & Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. Conference on Information and Knowledge Management. November 6−8, 2007, Lisboa, Portugal. Saatavilla verkosta osoitteessa <http://www.cs.umass.edu/~smucker/publications/smucker-statSig-cikm07.pdf> (käytetty 7.4.2008).

Strohman, T. et al. 2004. Indri: A language-model based search engine for complex queries (extended version). Teoksessa Proceedings of the International Conference on Intelligence Analysis. Saatavilla verkosta osoitteessa: <http://ciir.cs.umass.edu/pubfiles/ir-407.pdf> (käytetty 24.2.2008).

Strzalkowski, T. et al. 1999. Evaluating natural language processing techniques in information retrieval. Teoksessa Tomek Strzalkowski (toim.) Natural language information retrieval. Dordrecht: Kluwer Academic Publishers. S. 113–145.

Svenska akademiens ordbok (SAOB). 2008. Stockholm: Svenska akademien. Saatavilla verkosta osoitteessa: <http://g3.spraakdata.gu.se/saob/> (käytetty 26.2.2008).

Svenska akademiens ordlista över svenska språket. 2006. 13. painos. Stockholm: Svenska akademien.

Vilkka, H. 2007. Tutki ja mittaa. Määrällisen tutkimuksen perusteet. Helsinki: Kustannusosakeyhtiö Tammi.

LIITE 1 ESIMERKIT HAKUAIHEISTA

CLEF2003-hakuaihe

<top>
<num> C142 </num>
<SV-title> Christo paketerar det tyska riksdagshuset </SV-title>
<SV-desc> Leta efter rapporter om konstnären Christos inslagning av det tyska riksdagshuset. </SV-desc>
<SV-narr> Det tog inslagningskonstnären Christo två veckor under juni 1995 att slå in hela det tyska riksdagshuset i Berlin. Leta efter rapporter om denna konsthändelse. Information om förberedelser eller genomförande är relevanta, liksom politiska debatter och beslut samt tekniska förberedelser. </SV-narr>
</top>

CLEF2002-hakuaihe

<top>
<num> C140 </num>
<SV-title> Mobiltelefoner </SV-title>
<SV-desc> Framtidsutsikter för användningen av mobiltelefoner. </SV-desc>
<SV-narr> Relevanta dokument tar upp framtidsutsikterna för användningen av mobiltelefoner samt utvecklingen av mobiltelefonindustrin. </SV-narr>
</top>

Ahlgren-hakuaihe

<top>
<num> C013</num>
<SV-title> Konferens om familjeplanering</SV-title>
<SV-desc> Vilka diskussioner fördes och vilka resolutioner antogs vid befolkningskonferensen om familjeplanering i Kairo?</SV-desc>
<SV-narr> Alla debattinlägg, förslag och resolutioner om familjeplanering från befolkningskonferensen är av intresse. Speciellt relevanta är ställningstaganden från olika länder, organisationer och grupper.</SV-narr>
</top>

LIITE 2 ESIMERKIT KYSELYISTÄ

RAKENTEISET KYSELYT

CLEF2003-kyselyt

Perusmuotoistettu, ei ositettu kysely hakuaiheesta 142:

```
<query>#combine(konstnär christos inslagning #syn(tyska tysk) riksdagshus)
</query>
```

Perusmuotoistettu, ositettu, ei eliminoitu kysely hakuaiheesta 142:

```
<query>#combine(konstnär christos inslagning #syn(tyska tysk)
#syn(riksdagshus #uw20(rik dag hus) #uw20(rik dags hus) #uw20(riks dag hus)
#uw20(riks dags hus) #uw20(riksdag hus) #uw20(riksdags hus)))
</query>
```

Perusmuotoistettu, ositettu, eliminoitu kysely hakuaiheesta 142:

```
<query>#combine(konstnär christos inslagning #syn(tyska tysk)
#syn(riksdagshus #uw20(riksdag hus) #uw20(riksdags hus)))
</query>
```

CLEF2002-kyselyt

Perusmuotoistettu, ei ositettu kysely hakuaiheesta 140:

```
<query>#combine(framtidsutsikt användning mobiltelefon) </query>
```

Perusmuotoistettu, ositettu, ei eliminoitu kysely hakuaiheesta 140:

```
<query>#combine(#syn(framtidsutsikt #uw20(fram tid utsikt) #uw20(fram tids
utsikt) #uw20(framtid utsikt) #uw20(utsikt framtids)) användning
#syn(mobiltelefon #uw20(bil mo telefon) #uw20(mobil telefon)))
</query>
```

Perusmuotoistettu, ositettu, eliminoitu kysely hakuaiheesta 140:

```
<query>#combine(#syn(framtidsutsikt #uw20(framtid utsikt) #uw20(framtids ut-
sikt)) användning #syn(mobiltelefon #uw20(telefon mobil)))</query>
```

Ahlgren-kyselyt

Perusmuotoistettu, ei ositettu kysely hakuaiheesta 13:

```
<query>#combine(föra resolution anta befolkningskonferens familjeplanering
kairo ) </query>
```

Perusmuotoistettu, ositettu, ei eliminoitu kysely hakuaiheesta 13:

```
<query>#combine(föra resolution anta #syn(befolkningskonferens
#uw20(befolkning konferens) #uw20(befolknings konferens))
```

```
#syn(familjeplanering #uw20(plan familj ring) #uw20(plane familj ring)
#uw20(planering familje)) kairo) </query>
```

Perusmuotoistettu, ositettu, eliminoitu kysely hakuaiheesta 13:

```
<query>#combine(föra resolution anta #syn(befolkningskonferens
#uw20(befolkning konferens) #uw20(befolknings konferens))
#syn(familjeplanering #uw20(familj planering) #uw20(familje planering)) kai-
ro) </query>
```

RAKENTEETTOMAT KYSELYT

CLEF2003-kyselyt

Perusmuotoistettu, ei ositettu kysely hakuaiheesta 142:

```
<query>#combine(leta rappporter konstnär christos inslagning tyska tysk riks-
dagshus) </query>
```

Perusmuotoistettu, ositettu, ei eliminoitu kysely hakuaiheesta 142:

```
<query>#combine(leta rappporter porter ort rapp konstnär christos inslagning
tyska tysk riksdagshus dag rik riksdag dags riks hus riksdags) </query>
```

Perusmuotoistettu, ositettu, eliminoitu kysely hakuaiheesta 142:

```
<query>#combine(leta rappporter konstnär christos inslagning tyska tysk riks-
dagshus riksdag hus riksdags)</query>
```

CLEF2002-kyselyt

Perusmuotoistettu, ei ositettu kysely hakuaiheesta 140:

```
<query>#combine(framtidsutsikt användning mobiltelefon) </query>
```

Perusmuotoistettu, ositettu, ei eliminoitu kysely hakuaiheesta 140:

```
<query>#combine(framtidsutsikt tid framtid tids utsikt framtids användning
mobiltelefon bil mo telefon mobil) </query>
```

Perusmuotoistettu, ositettu, eliminoitu kysely hakuaiheesta 140:

```
<query>#combine(framtidsutsikt framtid utsikt framtids användning mobiltele-
fon telefon mobil) </query>
```

Ahlgren-kyselyt

Perusmuotoistettu, ei ositettu kysely hakuaiheesta 13:

```
<query>#combine(föra resolution anta befolkningskonferens familjeplanering
kairo) </query>
```

Perusmuotoistettu, ositettu, ei eliminoitu kysely hakuaiheesta 13:

```
<query>#combine(föra resolution anta befolkningskonferens befolkning konfe-
rens befolknings familjeplanering plan familj ring plane planering familje
kairo) </query>
```

Perusmuotoistettu, ositettu, eliminoitu kysely hakuaiheesta 13:

```
<query>#combine(föra resolution anta befolkningskonferens befolkning konfe-
rens befolknings familjeplanering familj planering familje kairo)
</query>
```