Karsittuja ja perusmuotoisia kyselyitä ja hakemistoja käyttämällä saatujen tulosjoukkojen päällekkäisyys

Kirsti Kujala

Informaatiotutkimuksen pro gradu -tutkielma

Syyskuu 2007

Informaatiotutkimuksen laitos

Tampereen yliopisto

TAMPEREEN YLIOPISTO
Informaatiotutkimuksen laitos
KUJALA, KIRSTI: Karsittuja ja perusmuotoisia kyselyitä ja hakemistoja käyttämällä saatujen tulosjoukkojen päällekkäisyys
Pro gradu -tutkielma, 126 s., 14 liites.
Informaatiotutkimus
Syyskuu 2007

TIIVISTELMÄ

Tutkielman tarkoituksena oli muodostaa suomen- ja englanninkielisistä kyselyistä ja hakemistoista erilaiset versiot perusmuoto-ohjelmien ja karsinta-algoritmien avulla, jotta voitaisiin selvittää, missä määrin erilaisilla kyselyversioilla saadut tulosjoukot ovat keskenään päällekkäisiä. Suomenkielisessä aineistossa perusmuotoisten ja karsittujen kyselyversioiden lisäksi muodostettiin ositetut perusmuotoiset kyselyt. Aineistona tutkielmassa käytettiin suomenkielistä TUTKia ja englanninkielistä TREC-tietokantaa, jotka pitivät sisällään lähinnä sanomalehtiartikkeleja. Englanninkielisten kyselyiden perusmuotoistamiseen käytettiin perusmuoto-ohjelma Engtwolia ja suomenkielisten kyselyiden perusmuotoistamiseen Fintwolia. Englanninkielisten kyselyjen karsintaan käytettiin Porter-algoritmia ja suomenkieliset kyselyt karsittiin Snowball-ohjelmistolla. Tiedonhakujärjestelmänä oli osittaistäsmäyttävä Inquery.

Tutkielmassa tarkasteltiin päällekkäisyyden lisäksi myös kyselyjen tarkkuuksia. Englanninkielisessä aineistossa karsinta ja perusmuotoistaminen olivat tuloksellisuudeltaan hyvin samankaltaiset. Suomenkielisessä aineistossa ositettu perusmuotoinen ja perusmuotoinen kyselysarja olivat tuloksellisuudeltaan hyvin samankaltaiset. Sen sijaan perusmuotoistaminen ja karsinta, verrattiinpa karsitun kyselysarjan kanssa sitten ositettua perusmuotoista tai osittamatonta perusmuotoista kyselysarjaa, poikkesivat kahdella relevanssitasolla tuloksellisuudeltaan jopa niin paljon, että niiden väliltä löytyi käytännössä havaittavat erot. Erot johtuivat siitä, että karsinta oli tuloksellisuudeltaan heikoin näistä kolmesta kyselysarjasta.

Englanninkielisessä aineistossa perusmuotoisen ja karsitun kyselysarjan välinen päällekkäisyys oli melko suurta tarkasteltaessa päällekkäisyyttä kokonaisissa tulosjoukoissa, sillä se vaihteli 70 prosentista 74 prosenttiin. Kun päällekkäisyyden tarkastelu rajattiin TRECissa tulosjoukkojen relevantteihin osiin, päällekkäisyys vaihteli relevanssitasosta riippuen 39 prosentista 1 prosenttiin. Kun päällekkäisyyttä tarkasteltiin suomenkielisessä aineistossa kokonaisten tulosjoukkojen osalta, eniten päällekkäisyyttä oli ositetun perusmuotoisen ja perusmuotoisen kyselysarjan välillä (87–94 %). Toiseksi eniten päällekkäisyyttä oli perusmuotoistamisen ja karsinnan välillä (53–61 %). Vähäisintä päällekkäisyys oli suomenkielisen ositetun perusmuotoisen ja karsitun kyselysarjan välillä (47–57 %). Kun päällekkäisyyden tarkastelu rajattiin TUTKissa tulosjoukkojen relevantteihin osiin, päällekkäisyys vaihteli kolmen pareittaisen vertailun eri relevanssitasoilla 62 prosentista 1 prosenttiin. Suomenkielisessä aineistossa päällekkäisyyden määrän laskuun vaikutti keskeisesti karsittu kyselysarja. Päällekkäisyyden lasku pareittaisissa vertailuissa, joissa toisena osapuolena oli karsittu kyselysarja, johtui karsitun kyselysarjan heikommasta tuloksellisuudesta. Koska karsittu kyselysarja löysi relevantteja dokumentteja muita kyselysarjoja vähemmän, osoittautui niistä yhteisiksi verrattavan tulosjoukon kanssa vielä pienempi määrä.

Sisällysluettelo

1 Johdanto

Yksinkertaisimmillaan tietokantaan on tallennettu kustakin dokumentista vain yksi esitys eli vain yksi dokumenttiversio. Tavallisesti tiedonhaku tapahtuu siten, että yhtä tiedontarpeen esitystä eli kyselyä täsmäytetään dokumenttijoukkoon ja saadaan tulokseksi yksi tulosjoukko. Kyselyistä ja dokumenteista voidaan kuitenkin laatia useita erilaisia versioita. Näin tapahtuu esimerkiksi silloin, kun kaksi tiedonhakijaa laatii kyselyt saman tiedontarvekuvauksen pohjalta, sillä he eivät todennäköisesti valitse kyselyihinsä täysin samoja hakuavaimia. Näin saman tiedontarpeen pohjalta syntyy kaksi erilaista kyselyversiota. Erilaiset kyselyversiot voidaan siis muodostaa käyttämällä eri kyselyversioissa keskenään erilaisia ilmaisuja, jotka kuitenkin viittaavat samaan käsitteeseen. Tällöin eri kyselyversioiden sisältämät hakuavaimet poikkeavat toisistaan semanttisesti. Kun semanttisesti erilaisilla kyselyversioilla tehtyjä tiedonhakuja on verrattu toisiinsa, on niiden havaittu saavan sellaiset tulosjoukot, joissa ei ole juurikaan ollut toisilleen yhteisiä dokumentteja. Toisin sanoen semanttisesti erilaisten kyselyiden tulosjoukkojen päällekkäisyys on ollut vähäistä (mm. Katzer, McGill, Tessier, Frakes & DasGupta 1982, 266; McGill, Koll & Noreault 1979, 75–76).

Erilaiset versiot voidaan laatia myös muilla tavoin. Toisenlainen tapa on esimerkiksi kieliteknologian hyödyntäminen. Silloin kyselyistä ja dokumenteista laaditaan erilaiset versiot ohjelmallisesti, jolloin niiden sisältämät luonnollisen kielen sananmuodot käsitellään kunkin ohjelman periaatteiden ja käsittelysääntöjen mukaisesti. Tällaisia ohjelmia ovat esimerkiksi karsinta-algoritmit ja perusmuotoohjelmat. Tällöin eri versioissa käytetään samoja ilmaisuja, jotka kuitenkin poikkeavat toisistaan merkkijonotasolla siten, että ne ovat keskenään hieman erimuotoisia ja eripituisia. Täten eri versioissa käytetään samoja ilmaisuja ja eri versiot ovat keskenään semanttisesti samanlaisia. Tässä tutkielmassa käytetään ohjelmallisesti laadittuja kysely- ja dokumenttiversioita. Tutkielmassa on tarkoitus selvittää, esiintyykö tulosjoukkojen välillä vähän päällekkäisyyttä myös silloin, kun käytössä on erilaiset kieliteknologian avulla laaditut kysely- ja dokumenttiversiot, jotka ovat keskenään semanttisesti samanlaisia.

Tutkielmassa kyselyistä ja dokumenteista luodaan erilaiset versiot karsinnan, perusmuotoistamisen ja osituksen avulla, joskin englanninkielisessä aineistossa ositusta ei käytetä lainkaan. Aiemmissa perusmuotoistamista ja karsintaa käsitelleissä tutkimuksissa on tutkittu enimmäkseen sitä, mikä näistä menetelmistä parantaa tiedonhaun tuloksellisuutta eniten. Aikaisemmista tutkimuksista poiketen tar-

kastelussa ei tällä kertaa ole keskeisintä tuloksellisuus, vaan tulosjoukkojen päällekkäisyys. Tarkoituksena on tutkia, miten paljon tuota päällekkäisyyttä esiintyy, kun karsituilla kyselyillä tietokannan karsitusta hakemistosta saatuja tulosjoukkoja ja perusmuotoisilla kyselyillä perusmuotoisesta hakemistosta saatuja tulosjoukkoja verrataan keskenään. Lisäksi suomenkielisessä aineistossa on tarkoituksena tutkia, missä määrin tulosjoukot ovat päällekkäisiä, kun verrataan keskenään ositettujen ja osittamattomien kyselyversioiden tulosjoukkoja. Tässä työssä päällekkäisyydellä tarkoitetaan tulosjoukkojen yhteisten dokumenttien suhteellisia osuuksia.

Jotta päällekkäisyyden tutkimiseen saataisiin lisää syvyyttä, katsotaan, miten paljon päällekkäisyyttä on kokonaisissa tulosjoukoissa. Tällöin tarkastellaan, mikä osa verrattavien tulosjoukkojen kaikista dokumenteista, olivat ne sitten epärelevantteja tai relevantteja dokumentteja, osoittautuu kummallekin tulosjoukolle yhteiseksi. Toinen tapa on rajoittaa tarkastelu vain tulosjoukkojen relevantteihin osiin, jolloin tarkastellaan, miten suuri osa verrattavien tulosjoukkojen relevanteista dokumenteista osoittautuu molemmille tulosjoukoille yhteisiksi. Lisäksi on mielenkiintoista tarkastella päällekkäisyyttä tulosjoukon eri kohdissa, kuten tulosjoukon alku- ja loppupäässä, sen selvittämiseksi, vaihteleeko päällekkäisyyden määrä tulosjoukon eri kohdissa.

Päällekkäisyyttä tarkastellaan kolmella eri relevanssitasolla: liberaalilla, normaalilla ja tiukalla relevanssitasolla. Käytännössä tämä tarkoittaa, että liberaalilla relevanssitasolla tarkastellaan, missä laajuudessa tulosjoukkojen kaikki relevanteiksi arvioidut dokumentit, olivat ne sitten marginaalisesti relevantteja, relevantteja tai erittäin relevantteja, osoittautuvat verrattaville tulosjoukoille yhteisiksi. Normaalilla relevanssitasolla puolestaan katsotaan, miten paljon tulosjoukkojen relevantteja ja erittäin relevantteja dokumentteja löytyy molemmista tulosjoukoista. Tiukalla relevanssitasolla katsotaan, missä määrin tulosjoukkojen erittäin relevantit dokumentit osoittautuvat yhteisiksi kummallekin tulosjoukolle.

Päällekkäisyyden määrää selvitetään käyttämällä sekä suomen- että englanninkielistä aineistoa, sillä suomi on hyvä esimerkki morfologialtaan rikkaasta ja englanti morfologialtaan niukasta kielestä. Suomenkielisenä testikokoelmana käytetään TUTKia ja englanninkielisenä testikokoelmana on TRECin 7. ja 8. arviointikierroksen aineistot. Kumpikin testikokoelma koostuu suurimmaksi osaksi sanomalehtiartikkeleista. Käytettävien testikokoelmien dokumentit ja kyselyt ovat tekstiä, joten tut-

kimus on luonteeltaan tekstitiedonhakua. Lisäksi tutkimus toteutetaan laboratorioympäristössä, jossa testikokoelma tarjoaa vakioidun tutkimusympäristön ja -olosuhteet.

Työn alussa käsitellään lyhyesti tiedonhaun käsitteitä. Kolmas luku käsittelee luonnollisen kielen ominaisuuksia ja sitä miten luonnollinen kieli vaikuttaa tekstitiedonhakuun. Neljännessä luvussa käydään läpi erilaisia sananmuotojen käsittelytapoja ja sananmuotojen käsittelyä aiemmin käyttäneiden tutkimusten tuloksia. Viides luku keskittyy aiempiin päällekkäisyyttä tarkastelleisiin tutkimuksiin. Kuudennessa luvussa kuvataan tässä tutkielmassa käytettäviä aineistoja ja tutkimusmenetelmiä. Seitsemäs luku on puolestaan varattu tulosten esittämiseen. Kahdeksannessa luvussa käydään keskustelua ja yhdeksännessä esitetään johtopäätökset.

2 Tiedonhaku

2.1 Tiedonhaun tutkimus

Tiedonhaku on kiinnostunut prosesseista, jotka liittyvät tiedontarpeen kannalta relevantin tiedon esittämiseen, tallentamiseen, etsimiseen ja löytämiseen. Tiedonhaun tutkimuksen tarkoituksena on ymmärtää tiedonhaun prosesseja, jotta voidaan suunnitella, rakentaa ja testata tiedonhakujärjestelmiä, jotka mahdollistavat tiedon tehokkaan löytämisen ja välittämisen. (Ingwersen 1992, 49.) Ingwersen mainitsee määritelmässään muun muassa tiedonhakujärjestelmien testaamisen. Testaamisen kanssa käsi kädessä on tiedonhakujärjestelmien arviointi. Tiedonhaun yhtenä osa-alueena nähdään tiedonhaun- ja -tallennuksen evaluointi, joka ei ole kuitenkaan rajoittunut vain tiedonhakujärjestelmien testaamiseen ja arviointiin, vaan sen piiriin kuuluu paljon muitakin evaluoinnin kohteita. Tiedonhaun- ja -tallennuksen evaluoinnin tehtävänä on arvioida muun muassa tiedonhakuhakujärjestelmiä ja tietokantoja sekä tiedonhakujärjestelmien käyttöä ja tehtyjä tiedonhakuja (Alaterä & Halttunen 2003). Tiedonhaun evaluoivassa näkökulmassa keskitytään muun muassa tiedonhaun tuloksellisuuteen (Järvelin 1995, 33). Niinpä käsillä oleva työ sijoittuu näkökulmaltaan juuri tähän tiedonhaun evaluoivaan osa-alueeseen.

Tiedonhaun tutkimukseen liittyy keskeisesti relevanssin käsite. Relevanssi muun muassa ilmaisee tiedonhaun tehokkuuden arvioimisen kriteerit (Saracevic 1996, 202). Käsitteen keskeisestä asemasta huolimatta se on osoittautunut olevan vaikeasti määriteltävissä. Se, mitä relevanssilla tarkoitetaan, on toisaalta intuitiivisesti ymmärrettävissä, toisaalta on kyse niin monimutkaisesta asiasta, että relevanssista itsestään on tullut tutkimuksen kohde ja keskeinen informaatiotutkimuksen tutkimusala (Saracevic 1996, 215). Saracevic (1996) on tarkastellut useita tutkimuksia, jotka ovat sisältäneet erilaisia kuvauksia relevanssista. Hän esittää tämän tarkastelun tulokset erityyppisinä relevansseina.

Algoritminen relevanssi kuvaa tiedonhakujärjestelmän tiedonhaussa käyttämän algoritmin kykyä täsmäyttää kysely relevantteihin dokumentteihin. Algoritminen relevanssi toteutuu, mikäli algoritmi toimii tarkoituksensa mukaisesti. **Aiheenmukainen relevanssi** kuvaa kyselyn ilmaiseman aiheen ja löydettyjen dokumenttien aihealueen välistä suhdetta. Aiheenmukainen relevanssi toteutuu, kun löydetty dokumenti käsittelee samaa aihetta kuin kysely. **Kognitiivinen relevanssi** kuvaa käyttäjän tietämyksen tilan ja tiedontarpeen sekä löydettyjen dokumenttien välistä suhdetta. Kognitiivinen vastaavuus, informatiivisuus, uutuus sekä tiedon laatu ovat kriteereitä, joiden perusteella dokumentit arvioidaan kognitiivisesti relevanteiksi. **Tilannesidonnainen relevanssi** kuvaa tilanteen, tehtävän, käsillä olevan ongelman sekä löydettyjen dokumenttien välisen suhteen. Tilannesidonnainen relevanssi määräytyy sen mukaan, miten käyttökelpoinen dokumentti on päätöksenteossa, ongelmanratkaisussa ja epävarmuuden poistamisessa. **Affektiivinen relevanssi** kuvaa käyttäjän aikomusten, tavoitteiden ja motivaation sekä löydettyjen dokumenttien välistä suhdetta. Affektiivisen relevanssin tunnusmerkkejä ovat muun muassa dokumentin tai tiedon aikaansaama tyytyväisyys. (Saracevic 1996, 214.)

Vaikka relevanssi on näinkin monimutkainen ja monitahoinen ilmiö, on monissa tiedonhaun tutkimuksissa dokumentit arvioitu vain joko aiheeltaan relevanteiksi tai epärelevanteiksi. Tätä kutsutaan binääriseksi relevanssiasteikoksi. Relevanssi on operationalisoitu tiedonhakujärjestelmissä hyvin yksinkertaisella tavalla silloin, kun relevanssi on nähty binäärisenä. Sormusen (2002, 324) mukaan binääristen relevanssiarvioiden käyttöä on kritisoitu realismin puutteesta. Binääristä relevanssiasteikkoa käytettäessä relevantiksi arvioimiseen on saattunut riittää, että pieni osa löydetystä dokumentista on katsottu arviointiprosessin aikana relevantiksi riippumatta siitä, miten pieni tuo osa on ollut (TREC 2006). Tosiasiassa kaikki näin löysin kriteerein relevanteiksi osoittautuneet dokumentit eivät ole samanasteisesti relevantteja, eivätkä ne tyydytä hakijan tiedontarvetta samalla tavoin. Järkevämpää olisi käyttää moniportaisia relevanssiarvioita ja pyrkiä siihen, että tiedonhakujärjestelmät kykenisivät löy-

tämään erittäin relevantteja dokumentteja marginaalisesti relevanttien sijaan. (Sormunen 2002, 324–329.) Niinpä joissakin tutkimuksissa on alettu käyttää moniportaisia relevanssiarvioita. Tässä työssä relevanssi nähdään aiheenmukaisena relevanssina ja käytössä olevat testikokoelmat sisältävät moniportaiset relevanssiarviot.

Relevanssin käsite liittyy myös tiedonhaun tulosten tarkasteluun. Yleensä tiedonhaun tuloksia mitataan ja esitetään saannin ja tarkkuuden avulla. **Tarkkuus** kuvaa tulosjoukon relevanttien dokumenttien suhdetta kaikkiin kyselyllä löydettyihin dokumentteihin. **Saanti** kuvaa tulosjoukon relevanttien dokumenttien suhdetta tietokannan kaikkiin relevantteihin dokumentteihin. Tarkkuutta tarkastellaan yleensä eri saantitasoilla. Esimerkiksi 10 % saantitaso tarkoittaa, että tietokannan relevanteista dokumenteista on löydetty yksi kymmenesosa.

Tässä työssä **dokumentilla** tarkoitetaan tekstidokumentteja, olivatpa ne sitten kokotekstidokumentteja tai bibliografisia viitteitä sellaisiin.

2.2 Tiedonhaun tasoperiaate

Sitä osaa ihmisen tietämyksestä tai informaatiosta, jota ollaan aikeissa kommunikoida muille, voidaan tarkastella käsiterakenteena. Käsiterakenne koostuu käsitteistä ja niiden välisistä suhteista. Jotta tämä käsiterakenne voidaan kommunikoida eteenpäin, sen pitää tulla konkreettisesti ilmaistuksi jonkinlaisten ilmaisujen avulla esimerkiksi puheen, tekstidokumentin, piirroksen tai muun sellaisen muodossa. (Järvelin 1995, 68.) Koska tutkielmassa on kyse tekstitiedonhaun tutkimisesta, keskitytään tässä tekstidokumenttien muodossa tapahtuvaan kommunikointiin. Jos tekstidokumentin kommunikoinnissa käytetään apuna tietokonetta, käsittelee tietokone näitä tekstidokumenttien ilmaisuja pelkkinä merkkijonoina. Tämän pohjalta voidaan nimetä kolme tiedonhaun tasoa: **käsitetaso**, **ilmaisutaso** ja **esiintymätaso**. Kyseiset tasot voidaan tunnistaa äskeisestä tiedon kommunikoijan tai tiedontuottajan toiminnan kuvauksesta, mutta myös seuraavasta tiedontarvitsijan toiminnan kuvauksesta. Tietoa tarvitessaan hakija tiedostaa tiedontarpeensa ja pyrkii määrittelemään tuon käsiterakenteessaan olevan aukon. Ilmaisutasolla hakija ilmaisee tiedontarpeensa yleisluontoisesti hakupyyntönä. Joko hän itse tai välittäjä pohtii sopivia ilmaisuja eli ilmaisutason hakuavaimia, joilla hakupyyntö tulisi hyvin kuvatuksi. Nämä hakuavaimet syötetään hakujärjestelmälle, joka käsittelee laadittua kyselyä pelkkinä merkkijonoina. (Järvelin 1995, 68–72.)

Vaikka hakujärjestelmälle voidaan syöttää kyselynä täsmälleen samat ilmaisut, joita käytettiin ilmaisutasolla, tulee ilmaisutaso ja esiintymätaso nähdä toisistaan erillisinä ja erilaisina. Esiintymätason merkkijonot toki edustavat ja heijastelevat ilmaisutason ilmaisuja, mutta ne eivät välttämättä vastaa toisiaan sataprosenttisesti, sillä ilmaisutason hakuavaimet voidaan esikäsitellä vaikkapa karsinta-algoritmin avulla ennen niiden syöttämistä hakujärjestelmälle, jolloin hakuavainten ja merkkijonojen välinen vastaavuus ei ole täydellinen. (Järvelin 1995, 72–73.) Toisaalta esiintymätason merkkijonojen ja ilmaisutason hakuavainten suhde on niin läheinen, että myös esiintymätasolla puhutaan usein hakuavaimista (Järvelin 1995, 187). Myös tässä työssä käytetään hakuavain nimitystä myös kyselyn esiintymätason merkkijonoista puhuttaessa.

Lisäksi selvennetään nimitysten hakuavain, hakusana ja hakutermi käyttöä. Käytettäessä kyselyissä luonnollisen kielen yksittäisiä sanoja ja yhdyssanoja on sopivaa käyttää nimitystä **hakusana**. **Hakutermi** taas on dokumentaatiokielen tai muun erikoiskielen termi, joka poikkeaa luonnollisen kielen sanasta sillä, että sen merkitys on rajatumpi ja täsmällisempi, mikä tekee siitä luonnollista sanaa yksiselitteisemmän. **Hakuavain** on yleisnimitys, joka kattaa hakusanat, hakutermit sekä yleiskieleen sisältyvät lyhenteet kuten (SPR) ja koodit esimerkiksi 20 °. (Järvelin 1995, 175–177.) Niinpä tässä työssä käytetään nimitystä hakuavain sen yleisluonteisuuden vuoksi.

2.3 Tietokanta ja sen rakenne

Tietokanta on jotakin käyttötarkoitusta varten laadittu kokoelma, joka sisältää toisiinsa liittyviä tietoja, jotka on tarpeen säilyttää (Laine 2005). Tietokanta koostuu useimmiten useista tiedostoista. **Tiedostot** koostuvat puolestaan useista tietueista. **Tietue** kuvaa määrämuotoisella tavalla jotain kohdetta tai kohteiden välisiä suhteita. Esimerkiksi opiskelija voi olla tällainen kohde, jolloin kyseistä opiskelijaa kuvaavia tietoja voidaan tallentaa opintorekisteriin tietue per opiskelija periaatteella. Tietue voi myös sisältää kuvattavan kohteen kokonaisuudessaan, esimerkiksi artikkelin. Tietue voi olla joko rakenteeton tai rakenteinen sen mukaan jakaantuuko se tietoalkioihin. Rakenteinen tietue jakaantuu **kenttiin**, joista kukin sisältää yhden tietoalkion. **Tietoalkiot** ovat kuvattavien kohteiden tärkeäksi arvioituja ominaisuuksia (esimerkiksi nimi), suhteita tai tunnisteita (esimerkiksi opiskelijanumero). (Järvelin 1995, 11–12.)

Peräkkäistiedosto (sequential file, linear line) sisältää kaikki tietokannan tietueet valitun lajittelutavan mukaisessa järjestyksessä. Lajittelutapa voi olla esimerkiksi tietuenumero, jolloin tietueet on lajiteltu tietuenumeron mukaiseen numerojärjestykseen. Peräkkäistiedostoa käytetään yleensä vain tulostusvaiheessa eli tulosjoukkoa käyttäjälle esitettäessä, jolloin järjestelmä hakee tietueet käyttäjän nähtäväksi peräkkäistiedostosta. (Järvelin 1995, 95–96; 101–104.)

Lähes kaikki tekstitiedonhakujärjestelmät käyttävät käänteistiedostorakennetta. **Käänteistiedosto** (inverted file) muodostetaan poimimalla kenttien sisältämät merkkijonot listaan, jossa niiden perään kirjataan sen tai niiden tietueiden järjestysnumerot, josta ne ovat peräisin. Nämä tietuenumerot toimivat peräkkäistiedostoon osoittavina osoitteina eli osoittimina. Kehittyneemmissä järjestelmissä osoitin ei ainoastaan ilmoita oikeata tietuetta, vaan ilmaisee tarkalleen, mistä kentästä ja mistä kohtaa kenttää merkkijono on peräisin. Mikäli merkkijono on esiintynyt useassa tietueessa, on sen perässä useita tietuenumeroita. Näin syntynyt merkkijonojen ja osoitinten lista järjestetään aakkosjärjestykseen. Käänteistiedostoa nimitetään usein myös tietokannan **hakemistoksi** tai **indeksiksi**. (Järvelin 1995, 96–97; 104.) Niinpä tässäkin työssä puhutaan enimmäkseen tietokannan hakemistosta.

Samasta peräkkäistiedostosta voidaan laatia useita käänteistiedostoja. Silloin joko kullekin kenttätyypille laaditaan oma käänteistiedosto (esimerkiksi nimekekentän merkkijonoille oma käänteistiedosto) tai luodaan usealle kenttätyypille yhteinen sekakäänteistiedosto. Sekakäänteistiedostossa merkkijonon perään liitetään tietuenumeron lisäksi tieto siitä, mistä kentästä se on poimittu. Käänteistiedoston tiedostorakenne poikkeaa peräkkäistiedoston rakenteesta siten, että se ei ole pelkkä merkkijonojen ja niiden osoitinten lista, vaan se on usein monitasoinen, puurakenteinen tai muuten omalla hakemistollaan varustettu, mikä tekee siitä hakujen kannalta peräkkäistiedostoa nopeamman. (Järvelin 1995, 97–98.)

Sanakirjatiedostoa (dictionary file) käytetään käänteistiedoston hakemistona. Se ilmoittaa kunkin merkkijonon perässä olevien osoitinten lukumäärän eli osoittaa hakuavaimen kirjausten määrän. (Järvelin 1995, 98.)

Tämä tutkimus perustuu karsitun, perusmuotoisen ja ositetun perusmuotoisen hakemiston käyttöön. **Karsittu hakemisto** on saatu aikaan käsittelemällä tietokannan sisältämien dokumenttien merkkijonot karsinta-algoritmilla. Merkkijonot ovat hakemistossa siis karsitussa muodossa. **Perusmuotoinen**

hakemisto on saatu aikaan muuttamalla tietokannan dokumenttien merkkijonot perusmuotoisiksi perusmuoto-ohjelman avulla. **Ositettu perusmuotoinen hakemisto** on ositettu versio perusmuotoisesta hakemistosta ja se sisältää alkuperäisten yhdyssanojen lisäksi yhdyssanojen yhdysosat. Karsinnasta, perusmuotoistamisesta ja osittamisesta kerrotaan lisää luvussa 4.

Tietokantaa rakennettaessa voidaan valita, viedäänkö käänteistiedostoon niin sanottuja **sulkusanoja**. Esimerkiksi tässä tutkielmassa käytettävistä tietokannoista sulkusanoiksi katsottavia avaimia ei ole poistettu käänteistiedostosta. Mikäli sulkusanoja ei poisteta käänteistiedostosta, on poisto järkevää tehdä kyselyille. Tiedonhaussa on toivottavaa, että sekä hakuavain että siihen täsmäytyvät merkkijonot ovat kyllin erottelukykyisiä eli kykenevät tehokkaasti jaottelemaan tietokannan dokumentit relevantteihin ja epärelevantteihin. Niinpä merkkijonot kuten prepositiot, artikkelit (a, an, the), konjunktiot, pronominit ja adverbit poistetaan kyselyistä sulkusanalistan avulla, sillä ne esiintyvät tietokannassa niin monta kertaa, ettei niiden käyttö hakuavaimina ole järkevää. Kettusen mukaan sulkusanoille on tyypillistä, että ne eivät kuvaa ulkoisen maailman olioita ja ilmiöitä, eivätkä siis kerro dokumentin aiheesta. Tiedonhaun kannalta keskeisimmät hakuavaimet ovat sanaluokaltaan substantiiveja (Kettunen 2005, 10–11; Salton & McGill 1983, 71–72).

3 Luonnollinen kieli

3.1 Kielen osajärjestelmät

Luonnollista kieltä voidaan luonnehtia monitasoiseksi, mutkikkaaksi ja joustavaksi järjestelmäksi. Kieli ei ole yksi jakamaton järjestelmä, vaan se sisältää osajärjestelmiä, jotka ovat keskenään monenlaisissa suhteissa. Kielen ja ennen kaikkea puhutun kielen kaksi keskeistä osajärjestelmää ovat äännejärjestelmä ja merkitysjärjestelmä, käytetäänhän kieltä merkitysten viestittämiseksi esimerkiksi puheen avulla. Merkitysten osajärjestelmää kutsutaan **semantiikaksi**, äännerakenteen osajärjestelmää **fonologiaksi**. Äänteiden ja merkitysten lisäksi kielessä on sanoja ja lauseita. Sanojen muodostamista ja sisäistä rakennetta tarkastelevaa osajärjestelmää sanotaan **morfologiaksi** eli muoto-opiksi ja lauseiden rakenteita tutkivaa osajärjestelmää **syntaksiksi** eli lauseopiksi. Vakiintuneiden sanojen osajärjestelmä on nimeltään **leksikko**. (Karlsson 1998, 12–15.) Käytännössä eri osajärjestelmien suhteet

ovat niin tiiviit, että ilmiöitä kuvattaessa käytetään yhdistettyjä käsitteitä kuten morfofonologia tai morfosyntaksi. Näistä osajärjestelmistä morfologia on tämän opinnäytetyön kannalta tärkein ja sitä tullaan käsittelemään lisää luvussa 3.3.

Kielen osajärjestelmä rakentuu sille tyypillisistä perusyksiköistä ja niiden välisistä suhteista. Fonologian perusyksiköitä ovat foneemit, leksikon yksiköitä itsenäiset sanat, sanavartalot tai päätetyypit, morfologian yksiköitä sanat ja morfeemit ja syntaksin yksiköitä lauseet, lausekkeet ja sanat. Merkitys syntyy vastaanottajan tulkitessa sanomaa käyttäen apuna sanomasta saamiaan erilaisia vihjeitä. Merkityksen abstraktiudesta johtuen on mahdotonta sanoa, mikä on merkityksen perusyksikkö. (Karlsson 1998, 15.)

3.2 Sana ja sanaan liittyvät käsitteet

Morfologian perusyksiköitä ovat sanat ja morfeemit, joista ensin määritellään käsite sana. Morfeemin käsitteenmäärittely tehdään jäljempänä. Sana on käsite, jota kielitieteen on ollut vaikea määritellä täsmällisesti, mutta joka on silti ymmärrettävä ja käyttökelpoinen. Sanasta esitettyjä määritelmiä on useita, mutta niistä on tarpeen esittää tässä yhteydessä vain tämän työn kannalta keskeisimmät.

Ensimmäisessä määritelmässä sana nähdään **lekseeminä** (lexeme) eli **leksikaalisena sanana**, joka edustaa kaikkia yhden sanan taivutusmuotoja samalla kertaa. Lekseemi on abstrakti yksikkö, sillä se ei voi esiintyä teksteissä sellaisenaan, vaan teksteissä esiintyy aina jokin lekseemin sananmuodoista. Suomen sananmuodot varvas, varpaan ja varpaillaan kuuluvat samaan sananmuotojen joukkoon eli ne ovat saman lekseemin esiintymiä. Käytännön sanakirjoissa lekseemin kaikkia sananmuotoja edustaa sanakirjamuoto eli ns. **leksikkomuoto**. Tästä syystä lekseemiä usein kutsutaan sanakirjasanaksi. Nominien sanakirjamuodoksi valitaan usein morfologisesti mahdollisimman yksinkertainen muoto. Verbeillä sanakirjamuotona on usein aikamuodon ja persoonan suhteen määrittelemätön muoto. Termiä **lemma** käytetään samassa merkityksessä kuin termiä lekseemi. (Häkkinen 1994, 138; Karlsson 1998, 187–188; Laaksonen & Lieko 2003, 29.)

Sama lekseemi voi esiintyä samassa tekstissä useita eri kertoja joko samassa tai eri taivutusmuodossa. Tämän pohjalta päästäänkin sanan muihin määritelmiin, sillä tekstiä tarkastellessa on tärkeää olla selvillä siitä, ollaanko laskemassa lekseemejä, **sananmuotoja** vai sanaesiintymiä. (Häkkinen 1994,

139.) Jotkin näistä sanaan liittyvistä käsitteistä ovat sellaisia, ettei niiden käyttö ja sisältö ole vakiintunut kielitieteessä (Karlsson 1998, 85–86). Seuraava jaottelu tekee kuitenkin niiden välille selkeän eron. Esimerkiksi virkkeessä Luen joko tuon lehden tai tämän lehden, mutta molempia lehtiä en lue esiintyy lekseemi lehti kolme kertaa, sananmuoto lehden kaksi kertaa ja sananmuoto lehtiä yhden kerran. Koko virkkeessä on kaksitoista sanaesiintymää, yksitoista eri sananmuotoa ja yhdeksän eri lekseemiä. Tekstissä olevien sanojen esiintymisfrekvenssiä tutkittaessa käytetään siis termiä **sanaesiintymä** tai **sane**, kun tarkoitetaan erikseen jokaista tekstissä esiintyvää sananmuotoa. (Häkkinen 1994, 139.)

Tässä työssä käytetään termejä sana, sananmuoto ja taivutusmuoto melko väljästi ja pitkälti toistensa synonyymeinä. Lekseemeistä ja sanaesiintymistä puhutaan mahdollisimman selkeästi ja kyseisiä nimityksiä käytetään silloin, kun on erityisen tärkeää selitettävän asian kannalta käyttää yksiselitteistä kieltä.

Siinä missä lekseemi on kaikkien taivutusmuotojen edustaja, **paradigma** muodostuu kaikista taivutusmuodoista. Sanan kaikki taivutusmuodot muodostavat yhdessä kyseisen sanan paradigman (Laaksonen & Lieko 2003, 61). Kielen morfosyntaktisista piirteistä riippuu, mitä taivutusmuotoja eri sanaluokkien sanat voivat saada. Näiden taivutusmuotojen sarja muodostaa lekseemin taivutusparadigman. Suomen substantiivien taivutusparadigmat muodostuvat 14 sijamuodon ja kahden luvun (yksikön ja monikon) avulla. Kielitieteessä ei olla yksimielisiä sijamuotojen määrästä. Jos kiistelty akkusatiivi katsotaan sijamuodoksi, on sijamuotojen määrä suomessa 15. Morfologian kannalta sitä ei kuitenkaan voida pitää sijamuotona, joten suomen substantiivien paradigmat muodostuvat 14 sijamuodon avulla. Myös adjektiivit saattavat taipua sijassa ja luvussa pääsanansa taivutuksen mukaisesti, esimerkiksi vanho+i+ssa autoissa. Tyypillisesti adjektiivit taipuvat vertailumuodoissa. Suomen verbien taivutusparadigmat muodostuvat puolestaan pääluokan, persoonan, luvun, tempuksen ja moduksen avulla. Pääluokka tarkoittaa verbien jaottelua kahteen pääluokkaan, aktiiviin ja passiiviin. Tarkempaa tietoa muun muassa tempuksesta ja moduksesta löytyy luvusta 3.4. Verbien ja nominien taivutusparadigmojen kokoon vaikuttaa se, että sanoja voidaan taivuttaa yhtä aikaa useammassakin taivutuskategoriassa, sillä esimerkiksi sananmuoto sano+isi+n on taivutettu sekä moduksessa että persoonassa. (Karlsson 1998, 106–111; Laaksonen & Lieko 2003, 77; 93.)

Englannissa sanaluokkien paradigmat jäävät pieniksi, koska asioita ei tyypillisesti ilmaista englannin kielessä morfologian avulla vaan erilaisin sanaston yhdistelmin. Esimerkiksi sijainti ilmaistaan englannissa usein preposition ja substantiivin yhdistelmällä eikä sanan perään liitettävien sijamuotojen avulla kuten suomessa. Myös omistusta ilmaistaan yhdessä erillisten possessiivipronominien ja substantiivien (esimerkiksi my dog ja your cat) avulla eikä omistusliitteiden avulla kuten sanoissa koirani ja kissasi. (Vannest, Bertram, Järvikivi & Niemi 2002, 84.) Englannin taivutuspäätteitä (inflectional affixes) on vain kahdeksan kappaletta: kaksi substantiiveihin liittyvää, kaksi adjektiiveihin ja adverbeihin liittyvää sekä neljä verbeihin liittyvää taivutuspäätettä (Klammer & Schulz 1992, 47–48). Niinpä englannin substantiivien taivutusparadigmat muodostuvat kahden luvun sekä kahden sijamuodon (nominatiivin ja genetiivin) avulla. Adjektiivien taivutusparadigmat puolestaan muodostuvat positiivin, komparatiivin ja superlatiivin avulla. Säännöllisten verbien taivutusparadigmat muodostuvat yksikön kolmannen persoonan, menneen aikamuodon, partisiipin preesensin ja partisiipin perfektin avulla. Epäsäännöllisten verbien taivutusparadigmat muodostuvat verbin yksikön kolmannen persoonan, verbin teeman 2. muodon, partisiipin preesensin ja verbin teeman 3. muodon avulla. (Klammer & Schulz 1992, 47–49.) Esimerkkejä englannin verbeistä löytyy luvusta 3.4. Loput englannin verbien ilmentymismuodot kuten monet aikamuodot, aikamuotojen kestomuodot ja passiivimuodot toteutetaan erilaisin apuverbin ja pääverbin yhdistelmin. Verbin paradigmalla tarkoitetaan tässä niin sanotun pääverbin taipumista eri muodoissa.

Taivutusparadigman koko eli paradigmaan kuuluvien sananmuotojen lukumäärä vaihtelee eri kielissä muutamasta sananmuodosta tuhansiin siitä riippuen, miten runsaasti kielessä on taivutuskategorioita (Karlsson 1998, 112). Samankin kielen sisällä eri sanaluokkien taivutusparadigmat ovat hyvin erikokoisia. Englanti on hyvä esimerkki kielestä, jossa paradigmaan kuuluvia sananmuotoja on vain muutamia. Englannin kielen verbeillä on nimittäin enintään viisi taivutusmuotoa: write, writes, wrote, written, writing (Mattila & Mattila 1997, 32). Englannin laskettavilla substantiiveilla voi olla neljä taivutusmuotoa: cruise, cruises, cruise's, cruices' (risteily). Englannin adjektiivit taipuvat kolmessa muodossa: tall, taller, the tallest. Suomen osalta kirjallisuudesta on löydettävissä muun muassa Koskenniemen (1985) ja Karlssonin (1983) esittämät kolmen sanaluokan paradigman kokoarviot. Tiivistetysti voidaan todeta, että täysparadigmaisella substantiivilla voi olla noin 2000 eri muotoa. Adjektiiveilla puolestaan on noin 6000 taivutusmuotoa, kun otetaan huomioon kaikki teoreettisesti mahdolliset yhdistelmät. Verbeillä eri muotojen kokonaismäärä on noin 12000. (Karlsson 1983, 356; Koskenniemi 1985, 20–21.) Koskenniemen (1985, 20–21) antamat arviot paradigmojen suuruudesta ovat

samankokoiset, mutta hän huomauttaa, etteivät annetut luvut pidä vielä sisällään johdoksia, joiden huomioon ottaminen nostaa luvut noin kymmenkertaisiksi. Myöskään englannin osalta lukemat eivät pidä sisällään johdoksia.

Vartaloa ja kantaa koskevat määritelmät eivät suomen ja englannin kielessä osu keskenään yksi yhteen. Niinpä kummankin osalta esitetään omat määritelmät. Ensin esitellään suomen kieleen soveltuvat määritelmät.

Vartalo on se sanan osa, joka jää jäljelle, kun taivutetusta sanasta erotetaan taivutuspäätteet. Vartalo ei välttämättä ole kaikissa paradigman muodoissa samanlainen. Kaikilla suomen sanoilla on vokaalivartalo, mutta konsonanttivartaloa niiltä ei välttämättä löydy. **Vokaalivartalolla** tarkoitetaan vokaaliin päättyvää vartaloa. Joskus sanan vokaalivartalo voi olla kahdenlainen kuten esimerkiksi sananmuodoilla **oppi**+vat ja **opi**+n. Sanan vartalossa on silloin astevaihtelun alainen konsonantti (esimerkiksi k, p, t tai kk, pp, ja tt), jonka vaihtelun johdosta sanalla on sekä **vahva** että **heikko vokaalivartalo**. Esimerkiksi sananmuodolla **katto** on vahva vokaalivartalo ja muodolla **kato**/n heikko vokaalivartalo. Joillakin sanoilla on lisäksi konsonanttiin loppuva **konsonanttivartalo**. Jos taivutuspäätteiden poiston jälkeen jäljelle jäävään osaan sisältyy jokin tunnus, puhutaan **alivartalosta**. Esimerkiksi puhui- on imperfektin alivartalo, sanoisi- konditionaalin alivartalo ja vanhoi- monikon alivartalo. (Laaksonen & Lieko 2003, 29–31.) Tiedonhaussa on usein puhuttu alivartaloiden sijaan **taivutusvartaloista**, joilla tarkoitetaan samaa kuin alivartaloilla. Suomen kielen sanoilla vartaloita voi olla perusmuoto mukaan lukien yhdestä viiteen kappaleeseen (Kettunen, Kunttu & Järvelin 2005, 477). Esimerkiksi kauppa-sanalla on viisi vartaloa: kauppa, kaupoi, kauppoi, kauppoj, kaupa. Sanan **kanta** on se johdetun sanan osa, joka jää jäljelle, kun taivutuspäätteet, tunnukset ja johtimet on erotettu (**onne**+ton, **onne**+ttom+uus) (Laaksonen & Lieko 2003, 113; Penttilä 2002, 127–128).

Jotta kyetään määrittelemään englannin käsitteet stem ja base, on syytä määritellä mitä tarkoitetaan juurella. Sanan osaa, joka ei jakaudu enää pienemmiksi osiksi, kutsutaan **juureksi** (Plag 2003, 11). Juurta ei siis voida analysoida enää pienempiin osiin. Juuri määritellään samalla tavoin myös suomessa. Englannissa käsitteiden **stem** ja **base** käyttö ei ole vakiintunutta. Toisinaan englannin sanojen rakenne esitetään niin, että puhutaan siitä sanan osasta, johon sekä johdin että taivutuspääte voidaan liittää. Yleensä tästä sanan osasta käytetään nimitystä base, mutta kyseisen nimityksen käyttö ei ole rajoittunut pelkästään tähän merkitykseen. Tämän määritelmän voitaisiin ajatella vastaavan suomen

kielen kannan määritelmää. Toisaalta englannin sanojen rakenteen kuvaamisessa voidaan painottaa taivutuspäätteiden lisäämistä ja puhua siitä sanan osasta, johon taivutuspäätteet lisätään. Silloin näkökulma on sanoissa, jotka ovat syntyneet johtamisen tuloksena. Usein tästä sanan osasta käytetään nimitystä stem, mutta myös muita nimityksiä on käytetty. Stem määritellään muodoksi, joka on morfologisesti kompleksinen eli moniosainen, sillä se muodostetaan yleensä johtamisen avulla juuresta. Juuri ja stem voivat olla samakin, sillä joskus taivutuspäätteet liitetään juureen, jonka seurauksena myös juuresta käytetään nimitystä stem. Koska tässä on kyse siitä sanan osasta, joka ei sisällä taivutuspäätteitä, voitaisiin ajatella tämän määritelmän vastaavan suomen kielen vartalon määritelmää. (ks. Lexicon of linguistics 2001; Matthews 1991; Plag 2003.)

3.3 Morfologia

Morfologia eli muoto-oppi on kielitieteen ala, joka tutkii sanojen rakenteita, muodostamista ja taivutusta sekä kielen morfeemeja. Morfologia selvittää, mitkä ovat tarkasteltavan kielen morfeemit, sekä miten morfeemeja voidaan liittää ja yhdistellä toisiinsa. Morfologia jaetaan yleensä taivutus- ja sananmuodostusoppiin. (Laaksonen & Lieko 2003, 27; Savolainen, Haakana, Lieko, Muikku-Werner & Mäntynen 1997.) **Taivutusoppi** käsittelee taivutukseen liittyviä asioita. Taivutuksessa käytetään morfologisia keinoja sananmuotojen, taivutusmuotojen muodostamiseksi lekseemistä. Näitä morfologisia keinoja ovat taivutuspäätteiden liittäminen vartaloon. (Häkkinen 1994, 125–128; Karlsson 1998, 106.) **Sananmuodostusoppi** tarkastelee sitä, miten olemassa olevan sanaston pohjalta voidaan muodostaa uusia sanoja. Keskeisimmät sananmuodostustavat ovat yhdistys ja johto. Yhdistyksessä yhdistetään kaksi tai useampi jo olemassa oleva sana toisiinsa, jolloin tuloksena on yhdyssana. Johdettaessa muodostetaan uusia sanoja muun muassa johdinten avulla. (Laaksonen & Lieko 2003, 109.)

Morfologian perusyksiköitä ovat sanat ja morfeemit, joista tässä määritellään käsite morfeemi. **Morfeemi** on kielen pienin yksikkö, jolla on merkitys tai kieliopillinen funktio. Morfeemeja on kahdenlaisia: vapaita ja sidonnaisia morfeemeja. Ainoastaan **vapaat morfeemit** voivat esiintyä yksinään. **Sidonnaiset morfeemit** yhdistyvät aina joko sanoihin tai toisiin morfeemeihin. Sidonnaisia morfeemeja nimitetään myös **affikseiksi**. Affikseja ovat **prefiksit, suffiksit ja infiksit**. Affiksit jaotellaan sen mukaan, miten ne sijoittuvat vartaloon nähden. Prefiksit sijoittuvat vartalon eteen ja suffiksit vartalon jälkeen. Vartalon sisäisiä affikseja kutsutaan infikseiksi. Karlssonin (1998, 101–102) mukaan

affikseista harvinaisimpia ovat infiksit, sillä niitä esiintyy vain muutamissa kielissä. Suffiksit ovat puolestaan yleisempiä kuin prefiksit (Karlsson 1998, 101–102).

Suomen kielessä on kuudenlaisia morfeemeja: vartaloita, johtimia, tunnuksia, taivutuspäätteitä ja kahdenlaisia liitteitä, nimittäin omistusliitteet ja liitepartikkelit, joista ainoastaan vartalot ovat vapaita morfeemeja (ks. Taulukko 1) (Laaksonen & Lieko 2003, 27). Englannin kielen morfeemeja ovat: vartalot, johtimet ja taivutuspäätteet. Englannin kielessä kaikki vartalot eivät ole vapaita morfeemeja vaan englannissa esiintyy myös **sidonnaisia vartaloita** (bound base / bound root) eli vartaloita, jotka voivat esiintyä vain yhdessä sidonnaisen morfeemin kanssa. Esimerkiksi circul- sanoissa circulate ja circulation on tällainen sidonnainen vartalo. (Klammer & Schulz 1992, 44; Plag 2003, 10.) Kummassakin kielessä prefiksit toimivat johtimina. Se on myös syy, miksi prefiksejä ei edellä lueteltu omana morfeemiryhmänään. Sen sijaan Taulukoissa 1 ja 2 prefiksit esitetään omana ryhmänään omalla paikallaan. Kyseiset taulukot havainnollistavat myös niitä tapauksia, jolloin vartalo nimityksen sijaan käytetään joitakin muita nimityksiä.

Morfeemi on käsite, jota tarkasti ottaen kuuluu käyttää kielen osajärjestelmien kontekstissa, sillä se on morfologian perusyksikkö. Todellisen kielen käytön eli puheen ja tekstin kontekstissa morfeemeita kutsutaan **morfeiksi**. Morfeemit siis toteutuvat puheessa ja tekstissä morfeina. Morfeista esiintyy variantteja. Tällaisia morfivariantteja tarkasteltavassa morfeemissa kutsutaan **allomorfeiksi**. Esimerkiksi suomen sijapäätteisiin lukeutuvalla inessiivimorfeemilla on kaksi variaatiota: ssa ja ssä. (Häkkinen 1994, 119; Laaksonen & Lieko 2003, 159–163.) Englannin sanoista sun ja sunny voidaan tunnistaa seuraavanlaiset morfit: sun, sunn+y. Morfeemi SUN voi siis toteutua kahtena morfina: sun ja sunn. (Creutz 2006, 13.) **Nollamorfista** puhutaan silloin, kun jokin morfeemin varianteista ei reaalistu ollenkaan. Esimerkiksi suomen imperatiivimuodossa tule! persoonapäätettä ei ole lainkaan. Persoonapäätteen allomorfina on tässä tapauksessa nollamorfi eli ei mitään. (Häkkinen 1994, 123.)

3.4 Morfeemien järjestys englannin ja suomen kielen sanoissa

Morfeemien järjestystä sanassa ja tätä järjestystä sääteleviä rajoituksia kutsutaan **morfotaksiksi** (Karlsson 1998, 103). Tässä luvussa tarkastellaan morfeemien järjestystä suomen ja englannin kielen sanoissa. Morfeemien järjestystä koskevat säännöt ovat kielikohtaisia. Taulukko 1 havainnollistaa morfeemien järjestystä suomen sanoissa ja Taulukko 2 morfeemien järjestystä englannin sanoissa.

TAULUKKO 1. Suomen kielen sanojen morfotaksi.

(Prefiksit/ johtimet)	Vartalo (Kanta ks. esimerkeistä kohta 3, s. 20)	Johtimet	Tunnukset: • Nominien luku (yks./mon.) • Passiivi • Tempus eli aikamuoto • Modus eli tapaluokka • Infinitiivien ja partisiippien tunnukset	Taivutuspäätteet: • Nominien sijapäätteet • Verbien persoonapäätteet	Liitteet: • Liitepartikkelit • Omistusliitteet eli possessiivisuffiksit

TAULUKKO 2. Englannin kielen sanojen morfotaksi.

Prefiksit/ johtimet	Vartalo (Juuri, kanta tai sidonnainen vartalo)	Johtimet	Taivutuspäätteet: • Verbin yksikön 3. persoonan –(e)s-pääte • Mennyt aikamuoto eli –ed-pääte tai epäsäännöllisen verbin teeman 2. muoto • Partisiipin preesensin tunnus eli –ing-pääte • Partisiipin perfektin tunnus eli –ed-pääte tai epäsäännöllisen verbin teeman 3. muoto • Substantiivin monikon –(e)s-pääte • Substantiivin genetiivimuoto (’s tai pelkkä ’) • Adjektiivin komparatiivin –er-pääte • Adjektiivin superlatiivin –est-pääte

Englannin kielen sanojen morfeemien järjestyksestä ei löytynyt kenenkään kielitieteilijän tekemää valmista esitystä, niinpä englannin morfeemien järjestys voitaisiin esittää myös esimerkiksi nimityksiä johdinprefiksit + vartalo / juuri / kanta / sidonnainen vartalo + johdinsuffiksit + taivutussuffiksit käyttäen. Englannin morfeemien järjestys tulee kuitenkin ymmärretyksi, käytettiin sitten näitä tai Taulukon 2 sisältämiä nimityksiä.

Taulukon 1 sisältöön liittyviä esimerkkejä:

1) Prefiksit, esimerkiksi **epä**+onni, **esi**+historia. Prefiksi on sanan alkuun liittyvä kielenaines, jonka lisääminen synnyttää uuden sanan. Taulukossa 1 prefiksit esitetään suluissa, koska suomi on kieli, jossa käytetään lähes pelkästään suffikseja.
2) Vartalot, esimerkiksi **talo**+ssa, **syö**+dä.
3) Johtimet, esimerkiksi ist+**ahta**+a, väitt+**ele**+n. Johdettaessa ei puhuta vartalosta vaan kannasta. Niinpä johdettaessa johtimet sijoittuvat lähimmäs kantaa. Johtimet voivat liittyä kannan sijasta myös toisiin johtimiin. Esimerkiksi verbissä heitt+**ele**+**ht**+**i**+ä on kolme johdinta.
4) Tunnuksia on viidenlaisia (ks. Taulukko 1). Tunnuksista kaikkein vierain lienee modusten luokka. Modukset liittyvät verbien taivutukseen. Modusten eli tapaluokkien avulla ilmaistaan subjektin suhtautumista predikaatin kuvaamaan toimintaan. Moduksia on neljänlaisia: indikatiivi, imperatiivi, konditionaali ja potentiaali. Muilla paitsi indikatiivilla on oma tunnuksensa, esimerkiksi konditionaalin tunnus on -isi. Niinpä indikatiivissa persoonapäätteet liitetään suoraan verbin vartaloon, esimerkiksi katso+**n**, syö+**tte** ja konditionaalissa konditionaalin tunnuksen perään, esimerkiksi katsoisi+**n**. Tunnusten alle sijoittuu myös nominien luku. Nomineita ovat substantiivit, adjektiivit, pronominit ja numeraalit. Esimerkki nominin monikon tunnuksesta löytyy sanasta koulu+**i**+ssa. Partisiippien tunnuksia ovat muun muassa –va, -vä, -nut, -nyt, -(t)tu ja –(t)ty, esimerkiksi sanoissa laula+**va**, melun+**nut** ja sano+**ttu**.
5) Taivutuspäätteet, esimerkiksi substantiivin inessiivipääte koulu+**ssa** ja verbin monikon 1. persoonan persoonapääte sano+**mme**. Sanaa päätteet käytetään tässä opinnäytteessä myös yleisnimityksenä ilman että nimenomaan tarkoitettaisiin taivutuspäätteitä.
6) Liitteet, esimerkiksi liitepartikkelit sanoi+**kin**, sait+**han** ja omistusliite talo+**nsa**.

(Karlsson 1998, 101–104; Korhonen, Vilkuna & Vihtari 2005; Laaksonen & Lieko 2003, 28; 93; 99; Savolainen ym. 1997.)

Seuraavassa esitetään vastaavia esimerkkejä englanniksi:

1) Prefiksit, esimerkiksi **pre**+history, **un**+happy, **fore**+tell. Englannin kielen prefiksejä ei käytetä taivutukseen.
2) Käsiteparit stem ja base sekä vartalo ja kanta. Esimerkiksi tarkasteltaessa monikkomuotoa disagreements, käytetään sanan osasta disagreement nimitystä stem. Huomaa, että kyseinen muoto on

johdos. Käsitteen stem määritelmän voitaisiin ajatella vastaavan suomen vartalon määritelmää. Juuresta muodostuu johtamisen avulla stem. Kun tarkastellaan taivutuspäätteen omaavaa sanaa fathers ja johtimen sisältävää sanaa fatherhood, käytetään substantiivista father nimitystä base. Käsitteen base määritelmän voitaisiin ajatella vastaavan suomen kannan määritelmää. Sidonnaiset juuret, esimerkiksi **later**+al. Sidonnaiset juuret ovat alkuperältään usein latinankielisiä.

3) Johtimet, esimerkiksi sananmuodoissa child+**ish**, child+**ly**, child+**hood**, child+**ish**+**ness**. Englannissa johdin (derivational affix) voi sijaita joko vartalon edessä tai jäljessä. Vartalon perässä oleva johdin muuttaa usein sanan sanaluokan. Johdin ei aina muuta sanan sanaluokkaa, mutta silloinkin sanan merkitys muuttuu.

4) Englannin taivutuspäätteitä (inflectional affixes) on vain kahdeksan kappaletta, joista neljä on verbien ja kaksi substantiivien taivutuspäätteitä. Loput kaksi ovat adjektiivien ja adverbien taivutuspäätteitä:

- verbin yksikön 3. persoonan -(e)s-pääte kuten sanoissa work**s**, pass**es**. Tätä päätettä käytetään verbin preesensmuodoissa.
- menneen aikamuodon –ed-pääte tai epäsäännöllisen verbin teeman 2. muoto, esimerkiksi watch**ed** ja **spoke**. Verbin mennyt aikamuoto tunnetaan imperfektinä.
- partisiipin preesensin –ing-pääte, esimerkiksi ask**ing**. Partisiipin preesensiä käytetään muodostettaessa aikamuotojen kestomuotoja. Partisiipin preesensillä on sekä verbin että substantiivin ominaisuuksia. Esimerkiksi lauseessa Smoking can be dangerous partisiipin preesens omaa substantiivin piirteitä.
- partisiipin perfektin –ed-pääte tai epäsäännöllisten verbien teeman 3. muoto, esimerkiksi ask**ed**, **shaken**. Partisiipin perfektiä käytetään yhdessä apuverbin kanssa perfektin, pluskvamperfektin, passiivin, 2. futuurin sekä 2. konditionaalin muodostamisessa.
- substantiivin monikkomuodon sisältämä –(e)s-pääte, jollainen esiintyy vaikkapa sanoissa book**s** ja dress**es**.
- substantiivin genetiivimuoto, joka on yksikössä muodossa girl’**s** tai monikossa muodossa taxpayers**’**. Taivutuspäätteiden määrää laskettaessa on substantiivin genetiivimuoto laskettu vain kerran. Monikon genetiivin heittomerkkiä ei siis ole laskettu omaksi päätteekseen.
- adjektiivien –er-pääte, joka esiintyy adjektiivien komparatiivimuodoissa, esimerkiksi dark**er**.
- adjektiivien –est-pääte, joka esiintyy adjektiivien superlatiivimuodoissa, esimerkiksi dark**est**. Komparatiivia ja superlatiivia ilmaisevat morfeemit esiintyvät myös pienessä määrässä adverbejä, esimerkiksi He drove long**er** and fast**er** than anyone else. Englannin komparatiivia ja

superlatiivia ilmaisevat morfeemit voidaan lisätä vain yksi- tai kaksitavuisten sanojen perään. (Klammer & Schulz 1992, 44–55; Lexicon of linguistics 2001; Mattila & Mattila 1997, 12–27; 78; 275.)

Englannin taivutuksessa affiksi voi sijaita vain vartalon lopussa. Mikäli sanassa on johdin, taivutuspäätteet liittyvät aina johtimen perään. (Plag 2003, 14–15.)

Kerrottakoon tässä yhteydessä vielä muutamia muita huomioita englannin substantiiveista ja verbeistä. Englannin substantiiveilla on aika paljon epäsäännöllisiä monikkomuotoja. Epäsäännöllisiä monikkomuotoja ei siis muodosteta yllä esitetyn monikon –(e)s-päätteen avulla, esimerkiksi a foot → feet. Lisättäessä edellä esiteltyjä päätteitä verbeihin saattaa sanassa tapahtua oikeinkirjoitusmuutoksia, esimerkiksi verbin invite lopussa oleva e katoaa, kun sanaan liitetään vokaalilla alkava –ing-pääte, jolloin tuloksena on muoto inviting (Mattila & Mattila 1997, 208). Lisäksi yksitavuisen ja yksivokaalisen sanan lopussa oleva konsonantti kahdentuu –ing-päätteen edellä, esimerkiksi split, splitting. Epäsäännöllisiä verbejä esiintyy englannissa melko runsaasti. Useimpien englannin epäsäännöllisten verbien mennyt aikamuoto muodostetaan joko vartalossa tapahtuvan vokaalinmuutoksen avulla, esimerkiksi run/ran, ride/rode tai nollamorfilla, jolloin preesens ja mennyt aikamuoto ovat kirjoitusasultaan samanlaisia, esimerkiksi cut/cut, hit/hit. Myös englannin epäsäännöllisten verbien partisiipin perfekti voidaan muodostaa vokaalin muutoksen tai nollamorfin avulla. Lisäksi epäsäännöllisten verbien partisiipin perfektien muodostaminen voidaan tehdä lisäämällä sanaan tavu –en, esimerkiksi eaten. (Klammer & Schulz 1992, 51; Mattila & Mattila 1997, 207.)

3.5 Johtaminen

Sanoja voidaan johtaa monella eri tavalla. Näistä eri johtamistavoista esitellään tässä vain yleisimmät. Jotkin johtamistavoista ovat yhteisiä suomen ja englannin kielelle ja jotkin esiintyvät vain toisessa kielessä. Johtamiselle on tyypillistä se, että johtamalla muodostettujen sanojen sanaluokka voi säilyä samana tai vaihtua toiseksi.

Edellisessä luvussa esitetyt esimerkit kuvasivat sekä taivutusta että johtamista. Ennen kuin johtamista tarkastellaan tässä luvussa syvällisemmin, selvitetään ensin, miten johtaminen ja taivutus eroavat toisistaan. Sekä taivutuksessa että johtamisessa sanoihin lisätään affikseja. Niinpä joskus voi olla vai-

kea hahmottaa niiden välinen ero. Taivutuksessa suffiksien lisääminen vartalon perään tuottaa sananmuotoja. Näin on esimerkiksi käytettäessä englannin verbin yksikön kolmannen persoonan -(e)s-päätettä tai substantiivin monikon -(e)s-päätettä. Johtaminen eroaa taivutuksesta siinä, että johdinsuffiksien tai johdinprefiksien lisääminen vartaloon synnyttää uusia lekseemejä. (Plag 2003, 14.)

Sanoja voidaan siis muodostaa johtamalla. Tavallisesti sanan johtaminen tapahtuu **johtimen avulla**, jolloin uusi sana muodostetaan liittämällä kantaan johdin. Johtamalla syntynyttä sanaa kutsutaan **johdokseksi** (derivative). Johtimia voi olla useita peräkkäin kuten esimerkiksi sanassa truth+**ful**+**ness**. Tämä johtamistapa esiintyy sekä suomen että englannin kielessä. Tässä yhteydessä voidaan tarkastella hieman myös johtimia. Lepäsmaan, Liekon ja Silverbergin (1996) mukaan johtimet pitävät sisällään useimmiten jonkin merkityksen, sillä ne ilmaisevat erilaisia asioita, esimerkiksi runsautta. Esimerkiksi johtimen -isa sisältävä sananmuoto kalaisa tarkoittaa sellaista, jossa on runsaasti kaloja. Tällaiset merkityksen omaavat johtimet luokitellaan **semanttisiksi johtimiksi**. Osa johtimista luokitellaan puolestaan kieliopillisiksi. **Kieliopillisen johtimen** funktiona on muuttaa sanan sanaluokka. Johtimet voivat olla merkitykseltään synonyymisia, homonyymisia ja polyseemisia. Niinpä johtimet voidaan luokitella äskeisen pääluokittelun lisäksi synonyymisiin, homonyymisiin ja polyseemisiin johtimiin. **Synonyymisista johtimista** on kyse, kun eri johtimilla on keskenään sama merkitys. **Homonyymiset johtimet** ovat keskenään samanmuotoisia, mutta niillä on eri merkitys. **Polyseemisilla johtimilla** on puolestaan monta merkitystä. (Lepäsmaa, Lieko & Silfverberg 1996, 20.)

Toinen suomen ja englannin kielessä käytetty johtamistapa on **takaperojohto** (back-formation), jossa muihin johtamistapoihin nähden toimitaan takaperoisesti luomalla johdoksen pohjalta uusi kantasana. Toisin sanoen sanoja johdetaan poistamalla suffiksi tai suffiksiksi tulkittu sanan osa. Esimerkkinä tästä ovat riehaantua → rieha, tarrata → tarra, editor → edit sekä self-destruction → self-destruct. (Laaksonen & Lieko 2003, 114; Plag 2003, 37.)

Suomessa johdoksia syntyy myös kielessä olevien **korrelaatiosuhteiden varassa**. Tällaisten johdosten kantasanaa on vaikeaa tai jopa mahdotonta osoittaa, sen sijaan näin lähekkäisten johdosten (esimerkiksi väsyä ja väsähtää), joista kumpikaan ei ole toisensa kantasana, sanotaan olevan samasta sanapesyeestä. Neljäs suomen kielen johtamistapa on **vartalon sisäinen johto**, jossa sanoja muodostetaan vaihtamalla vartalonsisäisiä äänteitä, esimerkiksi porskua ~ pärskyä. (Laaksonen & Lieko 2003, 113–114.)

Suomen kielessä sen tunnistaminen, milloin on kyse johtamisesta, voi olla vaikeaa, sillä pelkkä äänteellinen ja kirjoitusasun samanlaisuus tai sanojen välinen merkitysyhteys ei vielä ole osoitus johtamisesta. Esimerkiksi maja ei ole majakka-sanan kantasana, eikä selvä ole selkiää-sanan kantasana. (Laaksonen & Lieko 2003, 113–116.)

Englannissa sanoja voidaan johtaa myös lisäämättä affikseja. Sanoja voidaan muodostaa esimerkiksi muuttamalla substantiivi verbiksi lisäämättä vartaloon yhtään mitään. Esimerkiksi substantiivia water voidaan käyttää myös verbinä merkityksessä kastella kuten esimerkkilauseessa John waters his flowers every day. (Matthews 1991, 65; Plag 2003, 12; 107.) Tästä johtamistavasta on käytetty monia eri nimityksiä, mutta tyypillisesti sitä kutsutaan **konversioksi**.

3.6 Yhdyssanat ja niiden osittaminen

Johtamisen lisäksi sanoja voidaan muodostaa yhdistyksen avulla. Yhdyssana on helpointa tunnistaa silloin, kun se koostuu kahdesta toisiinsa liittyneestä sanasta kuten esimerkiksi yhdyssanassa lintulauta. Käytännössä kielet sisältävät tätä monimutkaisempia yhdyssanoja. Esimerkiksi englannin yhdyssana voi koostua useammasta kuin kahdesta osasta, sillä niissä voi olla neljä, viisi tai useampiakin osia kuten yhdyssanassa university teaching award committee member (Plag 2003, 133–135). Suomessa pitkähköjä yhdyssanoja ovat muun muassa syyttämättäjättämispäätös ja ammattikorkeakoulukokeilu. Yhdyssanan tunnistamisessa auttaa yhdyssanan määritelmä, jonka mukaan yhdyssana on kahdesta tai useammasta sanasta koostuva kokonaisuus, joka on kielen yksikkönä yksi sana ja myös edustaa tyypillisesti yhtä käsitettä (Korhonen ym. 2005).

Yhdyssanan osia kutsutaan **yhdysosiksi**. Yhdyssanat ovat joko alisteisia tai rinnasteisia sen mukaan, millainen yhdysosien välinen suhde on luonteeltaan. Yhdysosien välinen suhde on useimmiten alisteinen eli sellainen, jossa alkuosa määrittää ja kuvaa jälkiosaa tarkenteiden avulla, esimerkiksi **juhla**kenkä, **lenkki**kenkä, **dark**room ja **bath**room. Yhdyssanojen alkuosaa sanotaankin **määriteosaksi** (modifier) ja jälkiosaa **perusosaksi** (head). Yhdyssanojen toisessa tyypissä yhdysosien suhde on keskenään rinnasteinen, esimerkiksi kirjailija-kääntäjä, ruotsalais-suomalainen tai singer-songwriter. Rinnasteisia yhdyssanoja on kahdenlaisia, osa äskeisenlaisia ja osa sellaisia, jotka ovat erityisessä suhteessa niitä seuraavaan substantiiviin nähden, esimerkiksi the doctor-patient gap (Plag 2003, 146–

147). Yhdyssanan perusosa on usein hyödyllinen hakuavain, sillä se on usein koko yhdyssanan **hyperonyymi** eli yläkäsite.

Yhdyssanan perusosan perusteella määräytyy pitkälti koko yhdyssanan semanttiset ja syntaktiset ominaisuudet, joten jos perusosa on verbi myös yhdyssana on verbi, esimerkiksi deep-fry tai allekirjoittaa. Englannissa sama pätee myös tilanteessa, jossa perusosa on laskettavissa oleva substantiivi, sillä silloin myös yhdyssana on laskettavissa oleva substantiivi, esimerkiksi a beer bottle. Jos yhdyssana on monikkomuotoinen, esiintyy monikon tunnus perusosassa, esimerkiksi movie theater**s** tai elokuvateatteri**t**. (Plag 2003, 135.)

Kuten jo edellä olleista esimerkeistä on nähty, englannin yhdyssanojen osat kirjoitetaan usein erikseen (gross national product), mutta poikkeuksiakin löytyy, sillä joskus yhdyssana kirjoitetaan yhdysmerkkiä käyttäen (mother-in-law), tai kokonaan yhteen (database). On syytä huomata, että samankin yhdyssanan kirjoitustapa voi vaihdella, esimerkiksi teacup, tea-cup tai tea cup (Mattila & Mattila 1997, 275).

Eri kielet eroavat toisistaan siinä, miten paljon ne sisältävät yhdyssanoja. Suomelle tyypillistä on yhdyssanojen runsaus. Esimerkiksi Nykysuomen sanakirjassa yhdyssanoja ja yhdyssanojen johdoksia on noin 130 000 kappaletta eli yhdyssanojen osuus koko Nykysuomen sanakirjan sanamäärästä on 65 prosenttia (Saukkonen 1973, 337–338). Lepäsmaan ym. (1996, 12) mukaan yhdyssanoja on suomessa noin 44 % sanoista. Arviota englannin yhdyssanojen määrästä ei löytynyt.

Kun haetaan sellaisella kielellä kirjoitetusta tekstistä, joka sisältää paljon yhdyssanoja, on ongelmana, miten yhdyssanojen keski- ja loppuosat kyetään löytämään (Alkula 2000, 19). Tämän ongelman helpottamiseksi yhdyssanat voidaan osittaa, joskaan yhdyssanojen osittaminen ei aina johda tarkoituksenmukaiseen tulokseen. Yhdysosien merkitys ei välttämättä ole sama tai säily lähellä alkuperäisen yhdyssanan merkitystä. Niinpä yhdyssanojen ositus saattaa lisätä käänteistiedostoon dokumentin asiasisällöstä ohiampuvia sanoja. Tästä aiheutuvien haittojen minimoimiseksi on parempi sisällyttää hakemistoon alkuperäiset yhdyssanat ja lisätä yhdysosat niiden yhteyteen. (Hollink, Kamps, Monz & de Rijke 2004, 40.) Englannin erikseen kirjoitetut yhdyssanat on puolestaan kytkettävä toisiinsa tiedonhakua suoritettaessa esimerkiksi läheisyysoperaattorin avulla, jotta avaimet eivät täsmäytyisi muihin kuin yhdyssanoihin.

Joissakin kielissä yhdysosien välissä on yhdistävä elementti, joka voi olla yhden tai kahden kirjaimen pituinen esimerkiksi -s-, -e-, -en- ja niin edelleen. Tästä yhdistävästä elementistä on käytetty useita eri nimityksiä, esimerkiksi joining morpheme ja fogemorpheme pari mainitakseni. Suomenkielisenä nimityksenä voitaisiin kenties käyttää nimitystä **yhdysmorfeemi**. Esimerkiksi ruotsin sanassa skogsindustri yhdysosien skog ja industri välissä on yhdysmorfeemi s. Suomessa yhdysmorfeemeja ei esiinny, mutta ruotsissa ja saksassa niitä esiintyy runsaasti (Hollink ym. 2004, 40). Luonnollisestikaan kaikkia yhdyssanoja ei tällaisissa kielissä muodosteta yhdistävän elementin avulla niiden runsaasta esiintymisestä huolimatta. Englannin kielessäkin voi tavata yhdysmorfeemin, sellainen esiintyy esimerkiksi sanassa spokesman, mutta kaiken kaikkiaan niitä esiintyy hyvin harvoin (Krott, Baayen & Schreuder 2001; tässä Hollink ym. 2004, 40).

Jotta indeksitermit saisivat oikeanlaiset painot ja täsmäytyisivät tehokkaasti, yhdysmorfeemit on kyettävä poistamaan osittamisen yhteydessä ja saatava yhdysosille virheettömät perusmuodot. Yhdyssanojen ja yhdysmorfeemien vähäisen esiintyvyyden takia yhdysmorfeemien poistamisella ei ole vaikutusta englanninkieliseen tiedonhakuun. Eikä kyseistä seikkaa näin ollen tarvitse huomioida tämän tutkielman kyselyiden laadinnassa.

4 Tekstien sananmuotojen käsittely indeksointia varten

Kaikki kolme edellä esiteltyä morfologian ilmiötä taivutus, johtaminen ja yhdyssanat vaikuttavat tekstitiedonhakuun, sillä niistä aiheutuu ongelmia täsmäytykselle. Näiden ongelmien ratkaisemiseksi voidaan turvautua sananmuotojen käsittelyyn, jolla tarkoitetaan tässä sellaisia toimenpiteitä, joita dokumenttien ja kyselyiden merkkijonoille voidaan tehdä ennen niiden tallentamista hakemistoon tai ennen niiden syöttämistä hakujärjestelmälle täsmäyttämistä varten. Valittava käsittelymuoto vaikuttaa sekä käytettävän hakemiston että hakuavainten muotoon. Ensimmäinen valittavissa oleva vaihtoehto on jättää tietokannan hakemisto ja kyselyn hakuavaimet kokonaan käsittelemättä. Silloin sanat tallennetaan tietokannan hakemistoon siinä muodossa, jossa ne ovat esiintyneet dokumenttien teksteissä eli usein taipuneissa muodoissaan. Tällaista hakemistoa kutsutaan **taivutusmuotoiseksi hakemistoksi**. Jos hakemistoa ja kyselyä ei käsitellä mitenkään, joutuu käyttäjä sisällyttämään kyselyyn kaikki mahdolliset hakuavaimen taivutusmuodot saadakseen aikaan hyvän haun eli toisin sanoen löytääk-

seen kaikki kyseisen hakuavaimen sisältävät dokumentit. Tätä ongelmaa kyetään lievittämään sananmuotojen käsittelyn avulla. Taivutusmuotoisesta hakemistosta voidaan hakea taivutusmuotoisten hakuavainten lisäksi myös katkaistuja tai vartalomuotoisia hakuavaimia käyttäen, jos hakuavaimet ensin joko katkaistaan tai muutetaan vartalomuotoisiksi.

Tavanomaisin käsittelyn muoto on hakuavaimen oikealta puolelta tapahtuva **katkaisu**, jolloin sanan loppuosasta katkaistaan osa pois ja perään liitetään katkaisumerkki. Katkaistu hakuavain täsmäytyy kaikkiin samalla merkkijonolla alkaviin hakemiston sanoihin. Toivottavaa sananmuotojen käsittelyä voi olla myös katkaisu, joka tapahtuu hakuavaimen vasemmalta puolelta asettamalla katkaisumerkki sanan eteen, jolloin hakuavain täsmäytyy myös sanoihin, jotka ovat hakuavaimen alakäsitteitä. Tällöin alakäsitteitä ei tarvitse erikseen sisällyttää kyselyyn. Katkaisua käytettäessä hakemiston sanat ovat siis taivutusmuodoissaan ja vain kyselyn hakuavaimet katkaistaan. Kyselyn hakuavaimia voidaan käsitellä myös **vartalo-ohjelman** avulla, joka tuottaa hakuavaimesta sen vartalon tai vartalot, sillä yhden hakuavaimen käsittelyn tuloksena saatetaan saada useita vartaloita. Vartalo-ohjelman tuottamat vartalot ovat kieliopin kriteerit täyttäviä. Käytettäessä käsittelyn tuloksena saatavia vartaloita hakuavaimina pitää vartaloiden perään lisätä vielä katkaisumerkki, kun haetaan taivutusmuotoisesta hakemistosta täystäsmäyttävää järjestelmää käyttäen.

Karsinnan avulla samanmerkityksiset ja toisistaan vain hieman poikkeavat merkkijonot saadaan täsmäytymään toisiinsa. Karsinnassa sanan lopusta poistetaan suffikseja, jolloin sanasta jää jäljelle karsintavartalo (stem), joka täsmäytyy kaikkiin samalla merkkijonolla alkaviin sanoihin. Valittaessa sananmuotojen käsittelytavaksi karsinta suoritetaan karsinta sekä hakemistolle että hakuavaimille, jolloin karsitusta hakemistosta voidaan hakea vain karsittuja hakuavaimia käyttäen. Erona katkaisuun nähden on se, että karsinta tapahtuu automaattisin keinoin algoritmin avulla, joka käyttää sanojen morfologiset piirteet huomioivia sääntöjä karsintaa suorittaessaan. Katkaisu taas tapahtuu manuaalisesti käyttäjän oman harkinnan pohjalta. Vartaloiden tuottamisen ja karsinnan välinen ero on puolestaan se, että karsinnassa vartaloita syntyy vain yksi eikä syntyvä vartalo välttämättä ole kielitieteen näkökulmasta vartalo. Karsinnalla saatavaa vartaloa voitaisiinkin kutsua selkeyden vuoksi **karsintavartaloksi**. Vaikka karsinnan avulla syntyvä vartalo saattaa sattumalta olla muodoltaan samanlainen kuin kieliopillisesti oikeanlainen vartalo, ovat tuon vartalomuodon syntyä motivoineet tiedonhaun pyrkimykset, eivät kielitieteellisten kriteerien täyttäminen.

Neljäs luonnollisen kielen taipumisesta aiheutuvia ongelmia vähentävä käsittelyn muoto on **perusmuotoistaminen**. Perusmuotoistamisessa dokumenttien ja kyselyjen sanat palautetaan perusmuotoohjelman avulla perusmuotoonsa eli niin sanottuun sanakirjamuotoonsa, joka on suomen verbeillä I infinitiivi ja nomineilla yksikön nominatiivi. Englannin verbien perusmuotoa kutsutaan infinitiivin preesensiksi (McAlester ym. 1992, 51). Englannin muiden sanaluokkien perusmuodoista ei tiettävästi käytetä mitään erityistä nimitystä. Perusmuoto-ohjelmalla tehtävän käsittelyn tuloksena syntyvästä perusmuotohakemistosta haetaan perusmuotoisilla hakuavaimilla. Koska tämän työn kannalta keskeisimmät käsittelymuodot ovat karsinta ja perusmuotoistaminen, kerrotaan karsinnasta ja perusmuotoistamisesta laajemmin omissa alaluvuissaan ja esitellään muun muassa käsittelyssä käytettäviä ohjelmia yksityiskohtaisemmin. Sen jälkeen esitetään katsaus kyseisillä menetelmillä saaduista tutkimustuloksista.

Tyypillisesti hakemiston ja kyselyjen käsittely pyrkii luonnollisen kielen sanojen variaation vähentämiseen, mutta myös päinvastaista käsittelyä voidaan tehdä. Esimerkiksi **taivutusmuodot tuottavalla ohjelmalla** (full form generation) voidaan tuottaa sanan kaikki taivutusmuodot ohjelmallisesti. Haettaessa taivutusmuotoisesta hakemistosta kyselyyn saadaan varmuudella sisällytettyä kunkin hakuavaimen kaikki taivutusmuodot, esimerkiksi englannissa louse, louse's, lice ja lice's. Luonnollisesti tällainen käsittely soveltuu hyvin ainoastaan niihin kieliin, joilla on yksinkertainen morfologia eli sanat taipuvat vain muutamissa sijoissa.

Kyselyjen ja dokumenttien sananmuotojen käsittely on sidoksissa kulloiseenkin kieleen, sillä käytettävien algoritmien, vartalo-ohjelmien ja perusmuoto-ohjelmien tulee olla kullekin kielelle räätälöityjä. Tiedonhaun alalla on etsitty ja sovellettu myös kielestä riippumattomia käsittelymenetelmiä. Yksi tällainen menetelmä on erimittaisten **n-grammien** käyttö, jolloin kyselyjen ja dokumenttien ilmaisut esitetään esiintymätasolla n-grammeina. N-gram muotoisia hakuavaimia voidaan käyttää hakuavaimen kaikkien eri taivutusmuotojen hakemiseen. (Hollink ym. 2004, 42–43.) Sanat muutetaan n-grammeiksi yksinkertaisella tavalla. Ensin määritellään, monenko merkin pituisia n-grammien halutaan olevan. Sen jälkeen juoksevan tekstin sanat katkotaan tämän n:n pituisiksi. Kun ensimmäinen katkaisu on tehty, siirrytään eteenpäin yhden merkin verran ja suoritetaan uusi katkaisu. Esimerkiksi sanan hevonen kirjainkolmikot (tri-grammit) ovat hev, evo, von, one, nen. Kyselyn hakuavaimista muodostettuja n-grammeja verrataan hakemiston sisältämistä sanoista muodostettuihin n-grammeihin. Täsmäytys tapahtuu vertailemalla sanojen välillä olevien yhteisten n-grammien määrää.

Tuloksena palautetaan dokumentit, jotka ovat sisältäneet eniten kyselyn n-grammien kanssa yhteisiä n-grammeja.

Luonnollisessa kielessä esiintyvän variaation vähentämiseen tähtääviä käsittelyitä on perinteisesti käytetty tiedonhaun piirissä seuraavista syistä. Ensinnäkin käsittelyllä on pyritty pienentämään hakemiston kokoa ja vähentämään siten tallennuskapasiteetin tarvetta. Tämän merkitys käsittelyn motiivina on nykyään vähäisempi, koska tallennuskapasiteetin hankkiminen ei ole enää niin kallista. Toisekseen sananmuotojen variaatiota vähentävästä käsittelystä on se hyöty, että tiedonhaun tuloksellisuus paranee, oli kyseessä sitten morfologialtaan yksinkertainen tai rikas kieli. Tuloksellisuuden paraneminen perustuu muun muassa termifrekvensseihin. Termifrekvenssien vaikutusta havainnollistetaan tässä karsinnan avulla. Karsinnassa karsintavartalon termifrekvenssi lasketaan laskemalla yhteen kaikkien niiden sananmuotojen termifrekvenssit, jotka kyseisen karsintavartalon onnistuu kattaa. Karsintavartalon tilalle voidaan ajatella myös perusmuoto, joka kattaa kaikki saman sanan sananmuodot tai vartalo-ohjelman tuottama vartalo, joka kattaa kaikki saman sanan sananmuodot. Sananmuotojen käsittelyä käytettäessä termifrekvenssien laskeminen eroaa tavasta, jolla ne lasketaan taivutusmuotoisessa hakemistossa. Taivutusmuotoisessa hakemistossa kaikkia saman sanan eri taivutusmuotoja pidetään yksilöllisinä merkkijonoina ja erillisinä esiintyminä. Kunkin esiintymän termifrekvenssi ilmoitetaan erikseen eikä saman sanan eri esiintymien frekvenssejä lasketa yhteen, mikä luonnollisesti vaikuttaa lajitteluarvoon. Tuloksellisuuden paraneminen on edelleenkin tavoittelun arvoista. Kolmanneksi runsaasti taipuvissa kielissä kuten suomessa jonkinlainen hakuavainten käsittely on välttämätöntä, jotta haut ylipäätään onnistuisivat ja niillä saataisiin tyydyttävinä pidettäviä tuloksia.

Tässä luvussa esiteltiin käsittelymuodot yleisesti tuomalla esiin niiden väliset erot ja yhtäläisyydet. Seuraavissa alaluvuissa käsittelymuotoja ja niiden toteuttamisessa käytettäviä ohjelmia esitellään vielä yksityiskohtaisemmin.

4.1 Taivutusmuotoinen hakemisto ja vartalot

Kun tietokannan hakemistoa ei käsitellä mitenkään ennen tiedonhakua, on kyseessä taivutusmuotoinen hakemisto. Taivutusmuotoista hakemistoa käytettäessä kyselyjen hakuavaimet voivat olla taivutusmuotoisia, katkaistuja tai vartalomuotoisia riippuen siitä, mikä sananmuotojen käsittelytapa on valittu käyttöön.

Vartaloita on yleisesti ottaen käytetty tiedonhaussa harvoin. Vartaloiden tuottamista (stem generation, stem production) on nimitetty myös automaattiseksi katkaisuksi tai taivutusvartaloiden tuottamiseksi. Vartaloita tuottavista ohjelmista on käytetty nimityksiä taivutusvartalo-ohjelma tai vartalo-ohjelma. Vartaloita laaditaan siten, että vartalo-ohjelma tuottaa sille perusmuodossa syötetystä sanasta sen vartalot. Kielitieteen näkökulmasta vartalo-ohjelma tuottaa sekä vartaloita että alivartaloita. Vartalo-ohjelmat muodostavat sanan kaikki mahdolliset vartalot, jotka yhdessä kattavat sanan paradigman eri muodot. Suomen kielen sanat voivat saada 1–5 vartaloa, joten tuhansien teoreettisesti mahdollisten taivutusmuotojen määrä saadaan supistettua vartalo-ohjelmien avulla vain muutamaan vartaloon.

Lapsi → lapsi-, lapse-, last-

Kova → kova-, kovem-, kovi-

Yö → yö-, öi-

KUVIO 1. Esimerkkejä vartaloiden tuottamisesta.

Vartalo-ohjelmat toimivat pelkkien sääntöjen varassa, eikä niiden taustalla ole suurta sanastoa. Vartalo-ohjelmien käyttö edellyttää kahta asiaa. Ensinnäkin syötesanan tulee olla perusmuotoinen. Toisena edellytyksenä on, että käsiteltävän sanan sanaluokka tunnetaan. Yksinkertaisin tapa saada tieto hakuavaimen sanaluokasta on pyytää hakijaa ilmoittamaan se. Syötesanojen sanaluokat voidaan tunnistaa myös ohjelmallisesti lauseenjäsennysohjelman avulla. Yhdestä sanasta tuotetut vartalot muodostavat yhden kyselyn faseteista, jolloin ne ympäröidään sulkeilla ja yhdistetään toisiinsa TAI-operaattorilla. Kyselyä, jonka hakuavaimet ovat vartalomuotoisia, täsmäytetään taivutusmuotohakemistoon, jonka sanoja vartalo-ohjelmat eivät ole käsitelleet, vaan sanat ovat alkuperäisissä taipuneissa muodoissaan. Käytettäessä tällaisia vartaloita hakuavaimina täystäsmäyttävässä järjestelmässä niiden perään liitetään katkaisumerkki.

Koska osittaistäsmäyttävät järjestelmät eivät yleensä tue katkaisua, joudutaan katkaisu toteuttamaan jollakin toisella tavalla osittaistäsmäyttävää järjestelmää käytettäessä. **Katkaisun simulointia** ovat käyttäneet ainakin Kettunen, Kunttu ja Järvelin (2005), jotka sovelsivat katkaisun simulointia kahdella eri tavalla. Ensimmäisessä simulointitavassa kaikkia vartalo-ohjelman yhdelle syötesanalle antamia vartaloita verrataan tietokannan hakemistoon. Tällöin vartaloilla löydetään hakemistosta kaikki syö-

tesanan taipuneet muodot sekä muut syötesanan kanssa samanalkuiset sanat. Lopulliseen kyselyyn sisällytetään kaikki vartaloihin täsmänneet muodot. (Kettunen ym. 2005, 482–483.) Alkula (2000) on puolestaan ratkaissut katkaisuun liittyvän ongelman menettelyllä, jota hän kutsuu seulonnaksi. Seulonta ja Kettusen ym. (2005, 483) käyttämä toinen simulointitapa ovat keskenään hyvin samanlaisia. Tässä toisessa menettelytavassa varmistetaan, että lopulliseen kyselyyn hyväksytään vain sellaiset vartaloihin täsmäävät taivutusmuodot, jotka ovat syötesanan aitoja taivutusmuotoja. Tämä varmistus toteutetaan perusmuotoistamisen avulla. Liikkeelle lähdetään vartalo-ohjelman perusmuotoisesta syötesanasta. Myös tässä toisessa menettelytavassa katsotaan, mihin hakemiston sanoihin vartalo-ohjelman tuottamat vartalot täsmäävät. Sen jälkeen kukin vartalolla hakemistosta saatu sananmuoto syötetään perusmuoto-ohjelmalle, joka tuottaa vastaavat perusmuodot. Tämän jälkeen perusmuoto-ohjelman antamaa perusmuotoa verrataan alkuperäiseen perusmuotoiseen syötesanaan. Mikäli ne ovat täsmälleen samat, taivutusmuotohakemistosta saatu taivutusmuoto hyväksytään kyselyyn. (Alkula 2000, 117–118.)

4.2 Karsinta

4.2.1 Nimitysten ja algoritmien kirjo

Karsinnasta käytettävät nimitykset ovat vakiintumattomia ja kirjavia sekä englannin että suomen kielessä. Karsinnan tavanomaisimmat suomenkieliset nimitykset ovat stemmaus, karsinta ja typistäminen. Englannissa vastaavia nimityksiä ovat stemming, suffixing ja term conflation. Vaihtelua esiintyy myös sen suhteen, nimitetäänkö karsintaan käytettäviä algoritmeja stemmereiksi vai karsinta-algoritmeiksi. Nimityksiä stemmaus ja stemmeri on käytetty informaatiotutkimuksessa runsaasti ja ne on lainattu englannista muuttaen niitä hieman suomen kieleen sopivammiksi. Typistys, karsinta ja karsinta-algoritmi ovat nimityksiä, jotka ovat puhtaammin suomenkielisiä nimityksiä, mutta toisaalta niitä on käytetty harvemmin. Tässä työssä on päädytty jälkimmäiseen vaihtoehtoon eli nimitysten karsinta ja karsinta-algoritmi käyttämiseen ihan senkin perusteella, että juuri niitä nimityksiä myös Alkula (2000) on käyttänyt väitöskirjassaan. Nimitysten kirjon lisäksi karsinta-algoritmit ovat myös toimintaperiaatteiltaan kirjavia, sillä osa karsinta-algoritmeista on algoritmisia ja osa sanakirjaperustaisia. Tätä seikkaa käsitellään luvussa 4.2.3.

4.2.2 Karsinta ja siinä esiintyvät ongelmat

Karsinnassa semanttisesti läheisistä, mutta muodoltaan toisistaan poikkeavista sanoista poistetaan karsinta-algoritmin avulla suffikseja niin paljon, että ne saavat samanmuotoiset vartalot. Tämä on siis ainakin karsinnan pyrkimyksenä, vaikka, kuten seuraavasta kappaleesta nähdään, tämä tavoite ei aina toteudu käytännössä. Syötesanat annetaan karsinta-algoritmille taipuneissa muodoissaan. Kuten jo aiemmin on mainittu, ei jäljelle jäävä vartalo ole välttämättä kieliopin kriteerit täyttävä vartalo eikä edes oikea sana. Karsintavartaloa käytetään kyselyssä sellaisenaan ilman katkaisumerkkiä. Karsituilla hakuavaimilla haetaan karsitusta hakemistosta.

Paicen (1994, 42) mukaan karsinnassa tapahtuu väistämättä virheitä. Joko algoritmi karsii sanan lopusta liian vähän tai liian paljon eli tapahtuu ali- tai ylikarsintaa. **Alikarsinnassa** samaan käsitteeseen viittaavat sanat eivät typistykään yhtenäiseen vartalomuotoon. **Ylikarsinnassa** eri käsitteisiin viittaavat sanat saavat karsinnan tuloksena samanmuotoiset vartalot, vaikka niin ei kuuluisi tapahtua. Algoritmia suunniteltaessa joudutaan aina tasapainoilemaan alikarsinnan ja ylikarsinnan välillä. **Kevyt karsinta-algoritmi** karsii pääteaineksia maltillisesti välttääkseen ylikarsintaa, mutta osoittautuu sen seurauksena alikarsivaksi. Alikarsinnassa karsintavartalo jää niin pitkäksi, ettei se hakuavaimena käytettynä täsmäydy kaikkiin muodoltaan ja merkitykseltään läheisiin sanoihin. **Järeä karsinta-algoritmi** poistaa sanoista liikaa kaikenlaisia päätteitä, mikä heikentää tulosta useissa kyselyissä hakuavaimen täsmäytyessä epäsopiviin sanoihin ja palauttaessa epärelevantteja dokumentteja. Tällainen algoritmi on vahvasti ylikarsiva. Toisinaan järeä karsinta tuo kyselyyn tärkeän hakuavaimen, joka parantaa kyselyn tulosta paljon. (Hull 1996, 79–80; Paice 1994, 42.)

Porter (2001) määrittelee ali- ja ylikarsinnan lisäksi, mitä on väärin karsinta (mis-stemming). Hänen määritelmänsä täsmentää myös ylikarsinnan määritelmää. Ylikarsinnassa poistetaan todellinen pääte, jonka seurauksena eri merkityksen omaavat sanat saavat saman karsintavartalon. Väärin karsinnassa taasen sanasta poistetaan osa, joka näyttää päätteeltä, vaikka se tosiasiassa onkin osa karsintavartaloa. Esimerkiksi sanasta cheaply voidaan poistaa pääte –ly. Samanlaista poistoa ei voida kuitenkaan tehdä sanalle reply, koska siinä -ly ei ole pääte, vaikka se näyttääkin samanlaiselta kuin oikea pääte. Jos karsinta-algoritmi kuitenkin poistaa tuon osan reply sanasta, suorittaa se väärin karsinnan. Toinen esimerkki havainnollistaa ylikarsintaa. Tarkasteltavana on sekä sanaparin prove (todistaa) ja provable (todistettavissa oleva) että sanaparin probe (tutkia, ottaa selko) ja probable (todennäköinen, luultava)

karsinta. Ensinnäkin sanojen ulkoasun perusteella niiden voidaan olettaa olevan samaa alkuperää, ja testaaminen näyttäisi liittyvän jollakin tavalla sanojen merkitykseen. Kahden ensin mainitun sanan merkitykset ovat lähekkäiset, joten able-päätteen poisto saa aikaan toivotun tuloksen. Sen sijaan kahden jälkimmäisen sanan merkitykset eivät ole samanlaiset. Niinpä able-päätteen poistaminen sanasta probable olisi ylikarsintaa, sillä silloin poistettaisiin todellinen pääte, jonka seurauksena eri merkityksen omaavat sanat saisivat saman karsintavartalon. (Porter 2001.)

Väärin karsintaa ja ylikarsintaa voidaan pyrkiä välttämään käyttämällä karsinta-algoritmia, joka pitää sisällään sanakirjan. Sanakirja ei ratkaise näitä ongelmia kokonaan, mutta se voi toimia karsintaa parantavana välineenä. Luonnollinen kieli sisältää niin paljon poikkeuksia, ettei sanakirjaa käyttämälläkään voida poistaa kaikkia ongelmia. Sanakirjaa käyttävän karsinnan onnistumiseen vaikuttaa myös käytettävän sanakirjan laatu. Sanakirjan tulee olla kattava, ajantasainen sekä hyvillä sanan määritelmillä varustettu. (Porter 2001.)

Frakesin ja Foxin (2003, 29) havaintojen mukaan ylikarsintaa näyttäisi esiintyvän runsaasti sanoissa, joissa affiksit ympäröivät sanan keskellä olevaa vartaloa. Esimerkiksi sanassa ultranationalism vartalo on nation. Kuitenkin karsinnan myötä jäljelle jää pelkkä prefiksi, kun sana typistyy muotoon ultra. Tämän vuoksi he arvelivat karsinnan, jossa suoritetaan prefiksien poisto, osoittautuvan kannattavaksi. Silloin sanan vartalo tulisi paremmin esiin. (Frakes & Fox 2003, 29.) Asiasta on kuitenkin esitetty muitakin näkemyksiä. Hullin (1996, 82–83) mukaan prefiksien karsinta on epätoivottavaa, koska monien prefiksien kuten englannin anti-, un- ja il- prefiksien poistaminen muuttaa kyseisen sanan merkityksen täysin päinvastaiseksi. Myös Paice kertoo artikkelissaan, että prefiksien poistamisen on havaittu olevan hyödyllistä vain muutamilla tietyillä aloilla kuten lääketieteessä ja kemiassa. Sen sijaan suurinta osaa englannin suffikseista pidetään poistettavissa olevina. (Paice 1994, 42.) Siksi karsinta on tiedonhaun tutkimuksissa yleensä rajattu tarkoittamaan vain suffiksien poistamista.

Edellä kuvattujen karsintaan liittyvien ongelmien lisäksi on tarpeen tuoda esille vielä yksi ongelma. Karsinnan käyttämiseen morfologialtaan rikkaassa kielessä kuten suomessa liittyy sama ongelma kuin katkaisun käyttöönkin. Ensinnäkin taivutus on niin monimutkaista esimerkiksi tilanteissa, joissa saman sanan eri taivutusmuodoilla ei ole yhtään yhteistä kirjainta (yötä – öitä), ettei karsinta kykene yhden tuottamansa karsintavartalon avulla kuvaamaan tuota taipumista. Toisin sanoessa taivutuksessa esiintyy tilanteita, joissa karsinta ei kerta kaikkiaan onnistu tarkoituksenmukaisesti. Kuitenkin on-

gelmallisten tapausten määrä koko kokoelman sananmuodoista on niin vähäinen, että karsinta saattaa kuitenkin olla ihan käyttökelpoinen menetelmä. Karsinta-algoritmeja sofistikoituneempina ja soveliaampina välineinä runsaasti taipuvan kielen sananmuotojen käsittelyyn on pidetty perusmuoto-ohjelmia, sillä ne sisältävät kielen taivutusopin kuvauksen ohella suuren sanakirjan. Suomen käsittelyssä toisena riittävän sofistikoituneena käsittelytapana on pidetty vartaloiden tuottamista, sillä vartalo-ohjelmien tuottamat vartalot kattavat ja kuvaavat hyvin sanojen taivutuksen. Perusmuotoistamista tullaan käsittelemään lisää luvussa 4.3 ja vartaloista on kerrottu luvussa 4.1. Seuraavaksi käsitellään karsinta-algoritmeja.

4.2.3 Karsinta-algoritmit

Jo pelkästään englannin kieltä käsitteleviä algoritmeja on useita erilaisia. Keskeisin jaottelu tapahtuu yleensä jaottelemalla karsinta-algoritmit joko algoritmisiin tai sanakirjaperustaisiin (Porter 2001). **Algoritmiset karsinta-algoritmit** toimivat suffiksilistojen ja karsintasääntöjen varassa ilman sanakirjaa. **Sanakirjaperustaiset karsinta-algoritmit** nimensä mukaisesti käyttävät karsinnassa apuna suurta sanakirjaa. Lisäksi niiden sanoille suorittama analyysi on kielitieteellisesti motivoitunutta (Kettunen ym. 2005, 480–481). Tällaista suurta sanakirjaa käyttävää algoritmia voidaan kutsua myös **leksikaaliseksi karsinta-algoritmiksi** (lexical stemmer) (Kettunen ym. 2005, 491). Sanakirjan lisääminen karsinta-algoritmiin tekee siitä sofistikoituneemman sananmuotoja käsittelevän algoritmin, mutta myös hämärtää karsinnan ja perusmuotoistamisen rajaa, sillä perusmuoto-ohjelmakin käyttää suurta sanakirjaa. Porterin (2001) mukaan karsinnan algoritminen ja sanakirjaperustainen lähestymistapa eivät tosiasiassa ole erillisiä. Algoritminen karsinta-algoritmi voi sisältää pitkän listan poikkeuksista, jolloin nämä poikkeuslistat ovat tehokkaita minisanakirjoja. Sanakirjaperustaisen karsinta-algoritmin puolestaan tarvitsee yleensä poistaa vähintäänkin sanaa taivuttavat suffiksit, jotta algoritmin sisältämän sanakirjan käyttö olisi ylipäätään mahdollista. (Porter 2001.)

Englannin kielelle tarkoitetuista karsinta-algoritmeista kaikkein tunnetuimmat tiedonhaun tutkimuksen piirissä ovat S-algoritmi sekä Lovinsin ja Porterin algoritmit. Ne ovat ns. yleisalgoritmeja, joita ei ole laadittu mitään erityistä sanastoa silmällä pitäen, eikä minkään erityisen aineiston karsintaa varten. Jotkut algoritmit on voitu laatia jotakin tiettyä dokumenttikokoelmaa tai aihealaa varten, jolloin ne eivät ole yleiskäyttöisiä. Snowball-ohjelmisto on algoritmisten karsinta-algoritmien saralla uusi tuttavuus. Melko uutta on myös se, että karsinta-algoritmeja on saatavissa myös muiden kuin englan-

ninkielisten aineistojen käsittelyä varten. Hollink ym. (2004, 38) kertoo muun muassa Euroopan kielille laadituista karsinta-algoritmeista. Käyttöön on saatu karsinta-algoritmien kehittämistyökalu, jonka avulla on kyetty laatimaan karsinta-algoritmeja myös muille kuin englannin kielelle. Niinpä Snowball-ohjelmistosta on useita kieliversioita englanninkielisen version ohella. Snowball-ohjelmistot perustuvat osittain Porterin algoritmiin. (Hollink ym. 2004, 38.) Tunnetuimmat sanakirjaperustaiset algoritmit ovat Krovetzin (2000) karsinta-algoritmi sekä Braschlerin ja Ripplingerin (2003) saksan kielen käsittelyä varten kehittämät karsinta-algoritmit (Kettunen ym. 2005, 481).

Tässä työssä käytetään algoritmisia karsinta-algoritmeja, sillä englanninkielisten kyselyjen karsintaan käytetään Porter-algoritmia ja suomenkielisten kyselyjen karsintaan Snowball-ohjelmistoa. Niinpä seuraavaksi esitellään tunnetuimpien algoritmisten karsinta-algoritmien ominaisuuksia laajemmaltikin, mutta yleisluonteisesti. Heti sen perään esitetään tarkemmat kuvaukset tässä työssä käytettävistä algoritmeista. Algoritmiset karsinta-algoritmit toimivat toisistaan poikkeavilla tavoilla, mutta eroavaisuuksien lisäksi niistä on löydettävissä myös niille yhteisiä piirteitä ja ominaisuuksia.

Yleensä algoritmien karsintasäännöissä on määritelty alaraja, jota lyhyempiä karsintavartaloita ei voida tuottaa. Yleensä jäljelle jäävän karsintavartalon tulee jäädä vähintään kahden tai kolmen merkin pituiseksi. Karsintasäännöt ovat yleensä kolmiosaisia, jolloin sääntö ilmoittaa poistettavan päätteen, sisältää listan poikkeustapauksista ja kertoo toteutettavan toimenpiteen, joka on joko päätteen poisto tai kirjainmuutos. Kirjainmuutoksia on nimitetty myös uudelleen koodaussäännöiksi (recoding rules). Esimerkkisääntö: jos englannin kielen sana päättyy kirjaimiin ies (poistettava pääte), mutta ei kirjaimiin eies tai aies (poikkeustapaus) silloin tapahtuu kirjainmuutos ies → y (toimenpide), esimerkiksi studies → study. (Harman 1991, 8.)

Lisäksi eroja eri algoritmisissa karsinta-algoritmeissa on sen suhteen, kuinka monta erilaista poistettavaa suffiksia niiden suffiksilistasta löytyy. Esimerkiksi S-algoritmi poistaa sanojen lopuista vain monikon tunnuksen, kun taas Lovinsin algoritmi osaa poistaa jopa yli 260 erilaista suffiksia ja Porterin algoritmi noin 60. Kyseiset algoritmit ovat erilaisia myös siinä, toteutetaanko karsinta yksivaiheisena poistaen kerralla pisin poistettavissa oleva osa vai useammassa vaiheessa poistaen yksi suffiksi kerrallaan toistaen poistotoimenpidettä niin kauan kunnes sanasta ei enää löydy mitään poistettavaa. Eräs algoritmi tekee tarkistustoimenpiteen ennen lopullista karsintaa poikkeussanat sisältävän luette-

lon avulla, johon jäljelle jäävää vartaloa verrataan sen oikeellisuuden tarkistamiseksi. (Harman 1991, 8.)

Porter-algoritmi on karsintasäännöiltään kohtuullisen yksinkertainen ohjelma. Porter-algoritmissa on käytössä viisi erilaista karsintavaihetta. Mitä monimutkaisemman suffiksin käsiteltävä sana sisältää sitä useamman karsintavaiheen se joutuu läpikäymään. Ensimmäisessä vaiheessa poistetaan erityisesti monikon päätteitä ja partisiipin perfektit. Viidennessä vaiheessa ei enää poisteta varsinaisia suffikseja vaan tehdään pientä viimeistelyä. (Porter 1980, 131; 134–137.) Valitettavasti tietoa Snowball-ohjelmiston toimintaperiaatteista ei ollut tutkielman tekohetkellä saatavilla. Snowballin suomen kieltä karsivan ohjelmiston lähdekoodin perusteella Snowballia voidaan pitää kevyenä karsinta-algoritmina, sillä se ei pyri käsittelemään kaikkia kieliopillisia päätteitä ja tunnuksia (Kettunen 2005, 8).

Muun muassa Krovetz (1995; tässä Porter 2001) on väitöskirjassaan vertaillut algoritmisia ja sanakirjaperustaisia karsinta-algoritmeja keskenään. Hän havaitsi, että algoritmiset karsinta-algoritmit toimivat yllättävän hyvin sanakirjaperustaisiin karsinta-algoritmeihin verrattuna ja jäi pohtimaan syitä sen hyvään menestykseen (Krovetz 1995; tässä Porter 2001). Syitä ilmiölle esitti lopulta kuitenkin Porter (2001). Ennen syiden lähempää tarkastelua katsotaan lyhyesti Tomlinsonin (2003) tutkimusta, jossa algoritmisia ja sanakirjaperustaisia karsinta-algoritmeja on vertailtu muun muassa suomen osalta.

Myös Tomlinson (2003) on vertaillut algoritmisia ja sanakirjaperustaisia karsinta-algoritmeja keskenään. Suomen osalta parhaat tulokset saatiin leksikaalisella karsinta-algoritmilla algoritmisen karsinnan saadessa 13 prosenttiyksikköä huonompia tuloksia. Tutkimuksessa verrattiin leksikaalista SearchServer karsinta-algoritmia ja Snowball-ohjelmistoa keskenään. Kielinä olivat saksa, ranska, italia, espanja, hollanti, suomi, ruotsi, venäjä ja englanti. Sanakirjalla varustetulla karsinta-algoritmilla saavutettiin huomattavasti korkeammat keskimääräiset tarkkuusarvot kuin Snowball-ohjelmistolla, kun käsiteltävänä olivat suomen- ja saksankieliset kyselyt ja dokumentit. (Tomlinson 2003, 2–9.) Tomlinson (2003, 1) arvelee tämän johtuvan algoritmin kyvystä osittaa yhdyssanoja. Sen sijaan muiden kielten välillä erot eivät olleet tilastollisesti merkitseviä (Tomlinson 2003, 5).

Porter (2001) selittää, miksi algoritmiset karsinta-algoritmit toimivat niin hyvin, vaikka niitä käytettäessä tapahtuu paljon yli- ja alikarsintaa sekä väärin karsintaa. Hänen mukaansa alikarsinta on virhe, joka ei sinänsä heikennä tiedonhaun tuloksellisuutta. Vaikka sanat, joiden pitäisi saada sama karsin-

tavartalo, eivät sitä saakaan alikarsinnan seurauksena, ei haun tuloksellisuus huonone lähtötilanteeseen verrattuna, jos ei paranekaan. Enemmän haittaa koituu väärin karsinnasta, vaikka sekään ei käytännössä vaikuta tuloksellisuuteen, jollei sana sen seurauksena saa samaa karsintavartaloa eri merkityksen omaavan sanan kanssa. Niin ei useinkaan käy kuten seuraava esimerkki osoittaa. Esimerkiksi englannin -ate päätteen poistaminen voi olla toisinaan onnistunut ratkaisu kuten sanaparin luxury, luxuriate kohdalla, mutta se voi myös tuottaa karsintavartaloita, jotka eivät ole karsinnan jälkeen enää englanninkielisiä sanoja kuten sanantyngät enerv-ate ja accomod-ate osoittavat. Nämä lyhyenlaiset karsintavartalot eivät kuitenkaan saa samaa karsintavartaloa jonkin toisen sanan kanssa. Koska hakemiston sanat ovat samalla tavalla karsittuja täsmäävät nämä hakuavaimet vastaaviin hakemiston sanoihin, mutta eivät sanoihin, joiden kanssa niiden ei pitäisi täsmäytyä. Tiivistetysti voidaan sanoa, että alikarsinta ja väärin karsinta eivät vaikuta tiedonhaun tuloksellisuuteen kovinkaan paljon. Karsinta-algoritmien karsintavirheistä tiedonhaun tuloksellisuuteen vaikuttaa eniten ylikarsinta. Niinpä se, kumpi karsinta-algoritmityyppi suoriutuu paremmin ylikarsinnasta, on tuloksellisuudeltaan parempi. Sanakirjan käyttäminen karsinta-algoritmissa ei poista kaikkea ylikarsintaa, vaikka vähentääkin sen määrää. Tämä vähennys ei ole kuitenkaan niin merkittävä, että sanakirjaperusteisen karsinta-algoritmin tuloksellisuus osoittautuisi huomattavan paljon algoritmisen karsinta-algoritmin tuloksellisuutta paremmaksi. (Porter 2001.)

4.3 Perusmuotoistaminen

Perusmuotoon palauttamista on kutsuttu myös sananmuotojen normalisoinniksi ja morfologiseksi analyysiksi. Nimitysten suhteen saa kuitenkin olla tarkkana, sillä toisinaan normalisoinnilla ei tarkoiteta pelkästään perusmuotoistamista, vaan saatetaan sanoa karsinnan ja perusmuotoistamisen olevan kaksi sanojen taipumisen normalisointiin käytettävää menetelmää. Perusmuotoja tuottavista ohjelmista on käytetty muun muassa nimityksiä normalisoija, lemmatisoija, morfologinen ohjelma ja perusmuoto-ohjelma, joista tässä työssä käytetään viimeksi mainittua. Perusmuoto-ohjelmalle syötettävät sanat voivat olla taipuneissa muodoissaan. Niistä perusmuoto-ohjelma sitten tuottaa kunkin sanan perusmuodon. Perusmuotoisilla hakuavaimilla haetaan perusmuotohakemistosta, joten täsmäytys onnistuu ilman katkaisumerkkiä. Perusmuoto-ohjelmat sisältävät suuren sanakirjan ja kielen taivutussäännöt, joiden perusteella ne pystyvät päättelemään, mikä on jonkin sananmuodon perusmuoto. Lisäksi ne kykenevät osittamaan yhdyssanat yhdysosiinsa. Kehittyneet perusmuoto-ohjelmat voivat

perusmuotoon palauttamisen lisäksi ilmoittaa taipuneen sananmuodon sanaluokan ja sijamuodon. Näitä tietoja voidaan kutsua perusmuoto-ohjelman antamaksi luennaksi (reading).

Esimerkiksi:
Suunnitellaan → suunnitella
Maataloissa → maa#talo

Syöte: ilmaisin
Palaute ja luenta:
"ilmaisin" N NOM SG eli yksikön nominatiivi substantiivista ilmaisin
"ilmaista" V PAST ACT SG1 eli yksikön I persoonan imperfekti verbistä ilmaista
"ilmainen" A SUP NOM SG eli nominatiivi muotoinen superlatiivi adjektiivista ilmainen

Perusmuoto-ohjelman analysoimat syötesanat ovat kontekstistaan irrallaan olevia yksittäisiä sananmuotoja. Tällöin sananmuodot voivat olla monitulkintaisia. Monitulkintaisuus voi olla luonteeltaan joko **homonyymista** tai **polyseemista**. Homonymiassa kahdella tai useammalla sanalla on sama muoto sekä kirjoitus- että äänneasussa, mutta eri merkitys (Hakulinen & Ojanen 1993, 63). **Täydellisessä homonymiassa** kahden eri sanan paradigmat ovat täysin samanlaiset, jolloin ne ovat kaikissa muodoissaan keskenään identtisiä kuten vaara ’riski’ ja vaara ’kukkula’ (Penttilä 1975; tässä Laalo 1990, 18). Homonymia voidaan jakaa edelleen homofoniaan ja homografiaan. Homografiasssa vain erimerkityksisten sanojen kirjoitusasu on sama, kun taas homofoniassa vain erimerkityksisten sanojen ääntneasu on sama. **Taivutusmuotohomonymiasta** on kyse, kun kahdella tai useammalla sanalla on vain joitakin identtisiä taivutusmuotoja, esimerkiksi satoja voi olla joko sata- tai sato-sanan esiintymä. Näin on myös yllä esitetyssä perusmuoto-ohjelman luentaesimerkissä, jossa sananmuoto ilmaisin voi olla joko ilmainen, ilmaista tai ilmaisin sanan esiintymä. **Polysemiassa** yhdellä sanalla on useita merkityksiä eli se on monimerkityksinen. Esimerkiksi sanan kieli merkityksiä ovat ruumiinosa, puhuttu kieli sekä soittimen osa.

Kun perusmuoto-ohjelmalle syötetty sananmuoto on monitulkintainen, ohjelma palauttaa kaikkien niiden lekseemien perusmuodot, joiden kanssa analysoitavalla sanalla on yhteinen tai yhteisiä sananmuotoja. Perusmuoto-ohjelma ei kykene osoittamaan, minkä lekseemin perusmuodosta analysoitavassa sanassa on kyse. Se, mistä lekseemistä on kyse, on saatavissa selville vain, jos voidaan hyödyn-

tää syötesanan lauseen kontekstia ja kyseisestä tekstiyhteydestä saatavia vihjeitä. Perusmuoto-ohjelmat eivät kuitenkaan tee näin laajoja analyyseja. Ongelman ratkaisemiseksi pitäisi käyttää lauseenjäsennysohjelmaa, joka pystyy päättelemään lauseyhteyden perusteella, mikä analysoitavan sanan sanaluokka on kyseisessä lauseessa. Kun sanan sanaluokka on saatu selville, kyetään ratkaisemaan myös se, minkä lekseemin perusmuotoa voidaan pitää syötesanan oikeana perusmuotona.

Toisinaan perusmuoto-ohjelma ylitulkitsee sille annettuja syötesanoja tekemällä virheellisen tulkinnan yhdysosien suhteen (Järvelin & Kekäläinen 2002). Esimerkiksi seuraavissa syötesanoissa ohjelma näkee erikoisia yhdysosia:
Perusteluja → peruste # luja, perustelu
Ulkomailla → ulko # maa
Tällainen virheellisten tulkintojen tuottaminen heikentää hakujen tarkkuutta.

Perusmuoto-ohjelmassa käytettävän sanakirjan sisällön kattavuus vaikuttaa siihen, mitä sanoja perusmuoto-ohjelmat kykenevät tunnistamaan ja perusmuotoistamaan. Mikäli sanakirja ei sisällä jotakin sanaa, jätetään se alkuperäiseen taipuneeseen muotoonsa. Mikäli yhdyssanan yksikin osa on perusmuoto-ohjelmalle tuntematon, jää koko yhdyssana tunnistamatta (Alkula 2000, 106–107). Tunnistamatta jäämisen syynä on yleensä jokin seuraavista syistä: ohjelma ei tunnista sanaa kirjoitusvirheen takia, ohjelman sanakirja ei sisällä kyseistä joko vierasperäistä tai kotimaista erisnimeä tai sanakirjasta jostain syystä puuttuu jokin tarpeellinen yleisnimi. Kettunen (2005, 32–33) on tarkastellut perusmuoto-ohjelma Fintwolin sanakirjasta puuttuvien sananmuotojen ja sanaesiintymien määriä ja arvioinut niiden vaikutusta tiedonhakuun. Tutkitussa TUTK-tietokannan tekstiaineistossa oli Fintwolille tuntemattomia sananmuotoja 16,77 prosenttia kaikista sananmuodoista. Tuntemattomia sanaesiintymiä oli kaikista sanaesiintymistä 4,09 prosenttia. Tutkimuksessa esitettiin, että käyttäjän antamia hakuavaimia on pidettävä enemmän sanaesiintyminä kuin sananmuotoina, jolloin perusmuoto-ohjelman sanakirjasta puuttuvien sanojen vaikutus hakuihin asettunee lähemmäs sanaesiintymien prosenttilukua. (Kettunen (2005, 33–34.) Vastaavanlaista tietoa englannin kielen perusmuoto-ohjelma Engtwolin sanakirjasta ei ole saatavilla.

4.4 Aiempia sananmuotojen käsittelyä käyttäneitä tutkimuksia

4.4.1 Karsinta verrattuna taivutusmuotoiseen hakemiseen englanninkielisessä aineistossa

Harman (1987, 102–103; 1991, 7) vertaili tutkimuksessaan kolmea karsinta-algoritmia: S-algoritmia, Lovinsin algoritmia ja Porterin algoritmia. Tutkimuksen keskeisimmän testikokoelman muodosti Cranfield-kokoelman tiivistelmät ja otsikot. Lisäksi käytettiin Medlars- ja CACM-kokoelmia sen selvittämiseksi, onko testikokoelmalla tai testikokoelman kattamalla aihealueella vaikutuksia tuloksiin. Nämä kokoelmat olivat pienikokoisia, sillä niiden sisältämien dokumenttien määrät vaihtelivat 1033:sta 3204 dokumenttiin. Tiedonhakujärjestelmänä käytettiin osittaistäsmäyttävää IRX:ää. (Harman 1991, 8–9.)

Yksikään kolmesta karsinta-algoritmista ei parantanut tiedonhaun tuloksellisuutta merkittävästi. Vaikka karsinta ei parantanut tuloksellisuutta, ei se tarkoita, ettei karsinta olisi vaikuttanut hakuihin. Verrattaessa, monenko kyselyn kohdalla karsinta oli aiheuttanut parannusta tai huononnusta löydettyjen relevanttien dokumenttien määrällä tarkasteltuna joko kymmenen tai kolmenkymmenen dokumentin katkaisupisteen kohdalla, huomattiin että parantuneeseen tulokseen yltäneiden kyselyiden määrä oli lähes sama tai joskus jopa alhaisempi kuin huonontuneen tuloksen saaneiden kyselyiden määrä. Vaikka karsinta vaikutti myönteisesti yksittäisiin kyselyihin, ei parantuneen tuloksen saaneiden kyselyiden määrä ylittänyt selkeästi huonontuneen tuloksen saaneiden kyselyiden määrää. Kaikkein selkeimmin tämä kahtia jakaantuneisuus näkyi Lovinsin algoritmin kohdalla, sillä se tuotti kaikissa testikokoelmissa yleisesti ottaen eniten parannusta ja huononnusta kyselyiden tuloksiin. Porterin algoritmi sijoittuu tässä suhteessa lähemmäs Lovinsin algoritmia kuin S-algoritmia. (Harman 1991, 9–14.)

Hull (1996, 71–73) päätyi päinvastaisiin tuloksiin kuin Harman. Hän tutki viiden eri algoritmin: S-algoritmin, Lovinsin algoritmin muokatun version, Porterin algoritmin, Xeroxin taivutusmuotoalgoritmin ja Xeroxin johdosalgoritmin suorituksia. Johdosalgoritmi (derivational analyzer) palautti sananmuodot vartaloiksi tai kannoiksi ja poisti sanoista sekä suffikseja että prefiksejä. Johdosalgoritmin käsittelyn tuloksena saatiin aina oikeita englannin kielen sanoja. Taivutusmuotoalgoritmi (inflectional analyzer) palautti jokaisen sananmuodon sanakirjamuotoon. Algoritmeilla saatuja tuloksia verrattiin

tuloksiin, joissa karsintaa ei ollut käytetty. Testikokoelma sisälsi noin 180 000 Wall Street Journalin artikkelia ja tiedonhakujärjestelmänä oli vektorimalliin perustuva SMART. Käytössä oli 200 kyselyä, joista oli olemassa sekä pitkät että lyhyet kyselyversiot. Pitkinä kyselyversioina käytettiin TREC:in hakuaihekuvauksia sellaisenaan, jotka olivat lähes yhtä pitkiä kuin varsinaiset artikkelit. Vastapainoksi kyselyistä luotiin myös lyhyet versiot esittämällä hakuaihekuvauksen keskeiset seikat muutamalla lyhyellä fraasilla. (Hull 1996, 71–73.)

Hullin (1996, 73) mukaan kaikki algoritmit tuottivat 4–6 prosenttia parempia tuloksia verrattuna ilman karsintaa saatuihin tuloksiin, kun tuloksia tarkasteltiin laskemalla saanti ja tarkkuus kaikkien tasojen yli. Eri algoritmien väliset erot sen sijaan olivat hyvin pieniä. Tutkimuksessa myös arvosteltiin aiempia tiedonhaun tutkimuksia siitä, että suurin osa niistä oli lopettanut tulosten analysoinnin liian aikaisin ilman että ne olivat panostaneet yksityiskohtaisempaan analyysiin, esimerkiksi tekemällä tilastolliset analyysit ja ottamalla yksittäisiä kyselyitä yksityiskohtaisempaan tarkasteluun. (Hull 1996, 73.)

Hullin (1996, 75) mukaan hänen ja Harmanin saamat keskimääräiset saannin ja tarkkuuden tulokset olivat keskenään hyvin samankaltaiset. Vain johtopäätösten teko oli tapahtunut eri tavalla. Harman evaluoi karsinnasta koituvaa hyötyä 10 ja 30 dokumentin tarkastelun jälkeen. Hullin mukaan saannin ja tarkkuuden arvot voidaan mitata näin alhaisilla dokumenttien tarkastelutasoilla vain pienissä testikokoelmissa. Suurissa testikokoelmissa tarkoituksenmukaisemmat tarkastelutasot olisivat 10, 50, 100 ja 500. Suuressa testikokoelmassa keskimääräisen saannin laskeminen on merkityksetöntä alhaisilla dokumenttitasoilla, koska eri karsinta-algoritmien välillä olevia eroja ei kyetä osoittamaan tilanteessa, jossa kyselyiden tulosjoukot sisältävät satoja relevantteja dokumentteja. Tämä selittää sen, miksi Harman ei uskonut karsinnasta olevan hyötyä. (Hull 1996, 73–75.)

Kyselyiden yksityiskohtaisen analyysin perusteella Hull (1996, 76–83) teki seuraavat johtopäätökset:

1) Jonkin asteinen karsinta on aina hyödyllistä.
2) Taivutusmuoto-, johdos-, Lovinsin ja Porterin algoritmit ovat parempia kuin S-algoritmi, joka on kuitenkin parempi kuin karsimatta jättäminen. S-algoritmi on kuitenkin hyvin kilpailukykyinen muiden algoritmien kanssa silloin, kun tulosjoukon alusta tarkastellaan vain muutamaa dokumenttia. Neljän ensiksi mainitun algoritmin välillä ei ole eroja keskimääräisten tuloslukujen valossa.

3) Lovinsin algoritmi tuottaa muita algoritmeja huonompia tuloksia useissa kyselyissä, mutta toisaalta parantavaa muutamien kyselyiden tuloksia paljon enemmän kuin muut algoritmit. Näin on etenkin lyhyiden kyselyiden kohdalla. Mahdollinen selitys tälle ilmiölle on, että Lovinsin algoritmi on järeämpi kuin muut algoritmit.
4) Johdosalgoritmi toimii parhaiten, kun kyselyt ovat lyhyitä. Pitkien kyselyiden tulosjoukkoihin johdosalgoritmin käyttäminen tuo lisää epärelevantteja dokumentteja. Kun tuloksia tarkastellaan korkeilla saannin tasoilla, Porterin algoritmi toimii paremmin lyhyissä kuin pitkissä kyselyissä. (Hull 1996, 76–83.) Karsinta siis parantaa tuloksia, kun kyselyt ovat lyhyitä (Hull 1996, 72–76).

4.4.2 Sananmuotojen käsittely verrattuna taivutusmuotoiseen hakemiseen muissa kielissä

Karsintaa muiden kuin englanninkielisten dokumenttien ja kyselyiden sananmuotojen käsittelymuotona on tutkittu 1990-luvulta lähtien, jolloin karsintaa alettiin muutoinkin tutkia lähemmin. Syynä englannin dominanssiin sananmuotojen käsittelyssä voidaan pitää sitä, että karsinnan uskottiin soveltuvan hyvin vain morfologialtaan yksinkertaisiin kieliin, joiden taivutusta koskevat säännöt oli niiden vähäisen määrän takia helppo saattaa karsinta-algoritmin karsintasäännöiksi. Toisaalta englanti on vanhastaan ollut hallitseva kieli tiedonhaun tutkimuksessa. Englanninkielisen tutkimuksen rinnalle on viime aikoina noussut paljon muitakin tutkittavia kieliä ja myös suomenkielisen aineiston karsinnan tutkiminen liittyy tähän kehitykseen. Osoituksena tutkimuksen suuntautumisesta muihinkin kieliin voidaan pitää sitä, että suomen kielen käsittelyyn sopiva karsinta-algoritmi on ollut saatavilla nyt muutamia vuosia. Esimerkiksi Hollink ym. (2004) sekä Airio, Keskustalo, Hedlund ja Pirkola (2004) ovat käyttäneet kyseistä karsinta-algoritmia jäljempänä referoitavissa tutkimuksissa. Karsinnalla muissa kielissä saatujen tulosten lisäksi tässä luvussa selostetaan perusmuotoistamisella ja osittamisella saatuja tuloksia, kun tutkimusten aineistoina on käytetty muun kuin englanninkielisiä aineistoja.

Popovič ja Willett (1992, 384–389) kehittivät karsinta-algoritmin sloveeninkielisten kyselyiden ja dokumenttien käsittelyä varten. Heidän käyttämänsä testikokoelma sisälsi 504 dokumentin tiivistelmät ja tutkimus tehtiin osittaistäsmäyttävää tiedonhakujärjestelmää käyttäen. Koska sloveenin kieli on morfologialtaan paljon englantia rikkaampi, oli myös sitä käsittelemään laadittu karsinta-algoritmi englantia karsivia algoritmeja monimutkaisempi. Algoritmin suffiksilista sisälsi jopa 5276 karsittavaa

suffiksia. Tutkimuksessa käytettiin kolmenlaisia kyselyitä: karsittuja, taivutusmuotoisia sekä välittäjän (intermediary) manuaalisesti katkaisemia. Karsinnan hyötyä tarkasteltiin käyttämällä katkaisupisteenä kymmenennen dokumentin rajapyykkiä. Karsitut ja manuaalisesti katkaistut kyselyt olivat tuloksellisuudeltaan lähes yhtä hyvät. Sen sijaan erot karsittujen ja taivutusmuotoisten kyselyiden tuloksellisuudessa osoittautuivat suuriksi ja tilastollisesti merkitseviksi. Sen varmistamiseksi etteivät tutkimuksessa saadut tulokset johtuneet käytetystä testikokoelmasta, käännettiin testikokoelman dokumentit ja kyselyt englanniksi. Tällä kertaa vertailu tehtiin vain karsittujen ja taivutusmuotoisten kyselyiden välillä. Karsinta-algoritmina käytettiin tässä yhteydessä Porterin algoritmia. Taivutusmuotoiset englanninkieliset haut toimivat paremmin kuin taivutusmuotoiset sloveeninkieliset haut. Karsinta paransi englanninkielisten hakujen tuloksia vain hieman. Popovič ja Willett sanovat karsinnan olevan tehokasta, kunhan kieli, jonka käsittelyyn sitä käytetään, on morfologialtaan suhteellisen monimutkainen. (Popovič & Willett 1992, 384–389.)

Kunttu (2003) tutki, miten osittamattoman perusmuotohakemiston, ositetun perusmuotohakemiston ja taivutusmuotohakemiston käyttäminen vaikuttaa tiedonhaun tuloksellisuuteen todennäköisyyksiin perustuvassa tiedonhakujärjestelmässä. Testikokoelmana oli noin 54 000 suomenkielistä sanomalehtiartikkelia sisältävä TUTK ja tiedonhakujärjestelmänä osittaistäsmäyttävä Inquery. Ositetusta perusmuotohakemistosta saatiin paremmat tulokset kuin osittamattomasta tai taivutusmuotoisesta hakemistosta, joten ositettu perusmuotohakemisto oli tuloksellisuudeltaan kaikkein paras hakemisto. Näin ollen morfologisesti rikkaan ja paljon yhdyssanoja sisältävän kielen yhdyssanat kannattaa jakaa yhdysosiin ja tallentaa käänteistiedostoon ositettuina. (Kunttu 2003, 2; 72.)

Hollink ym. (2004) tutkivat sekä kielisidonnaisia että kieliriippumattomia menetelmiä kahdeksassa erikielisessä aineistossa. Tarkastelussa käytettiin yhtä 50 hakuaiheen sarjaa. Hakuaiheet kaikille kahdeksalle kielelle saatiin kääntämällä tuon sarjan hakuaiheet kullekin kielelle. Tosin suomenkielinen kokoelma kattoi vain 30 hakuaihetta. Heillä oli käytössään FlexIR niminen vektorimalliin perustuva tiedonhakujärjestelmä. Käytettyjä kielisidonnaisia menetelmiä olivat karsinta, perusmuotoistaminen ja yhdyssanojen osittaminen. Kieliriippumattomista menetelmistä käytössä olivat n-grammit. Kielisidonnaisia ja kieliriippumattomia menetelmiä tarkasteltiin ensin itsenäisinä menetelminä ja lopulta niitä myös yhdisteltiin keskenään esimerkiksi laatimalla n-grammeja karsituista sanoista. (Hollink ym. 2004, 34–46.) Kieliriippumattomia menetelmiä käsitellyttä osuutta ei käydä tässä läpi.

Tutkittuja kieliä olivat hollanti, englanti, ranska, saksa, italia, espanja, ruotsi ja suomi. Saaduista tuloksista referoidaan tässä vain suomen, ruotsin ja englannin tulokset. Hollink ym. (2004, 38) käyttivät näiden kielten käsittelyyn Snowball-ohjelmistoja. Verrattaessa karsittujen kyselyiden tuottamia tuloksia taivutusmuotoisten kyselyiden tuloksiin tehtiin seuraavanlaiset havainnot. Englanninkieliset kyselyt tuottivat hyvät tulokset jo kyselyjen kohdistuessa taivutusmuotohakemistoon. Näihin tuloksiin verrattuna karsinta paransi englanninkielisen tiedonhaun tuloksellisuutta vain hieman. Ruotsin kohdalla karsinta tuotti vertailukohtana käytettyihin taivutusmuotoisiin kyselyihin nähden paremman tuloksen, mutta parannus ei ollut merkitsevä. Sen sijaan suomen osalta karsinta paransi tuloksia merkitsevästi. Suomen kohdalla karsinta tuotti jopa kaikkein suurimman parannuksen kahdeksalla muunkielisellä aineistolla saadun parannuksen määrään verrattuna. (Hollink ym. 2004, 38–39.)

Kahdeksasta kielestä vain runsaasti yhdyssanoja sisältävien kielten (hollanti, suomi, saksa ja ruotsi) hakemistojen ja kyselyiden yhdyssanat ositettiin. Ensin selvitettiin, mikä on pelkän yhdyssanojen osittamisen vaikutus tuloksiin. Taivutusmuotohakemiston ja taivutusmuotoisten kyselyjen sisältämien yhdyssanojen osittaminen paransi tuloksia hollannin 4 prosentista suomen 18,7 prosenttiin, kun vertailukohtana käytettiin tuloksia, jotka oli saatu samaa hakemistoa käyttämällä ennen yhdyssanojen osittamista. Seuraavaksi tarkasteltiin, miten paljon yhdyssanojen osittaminen vaikuttaa karsitussa aineistossa. Karsitussa hakemistossa yhdyssanojen osittaminen tuotti 3,6–25,3 prosenttia parempia tuloksia, kun tuloksia verrattiin karsitulla hakemistolla ennen yhdyssanojen ositusta saatuihin tuloksiin. Parannus suomen osalta oli 9,8 prosenttia. (Hollink ym. 2004, 42.) Yhdyssanojen ositus ei ole siis käyttökelpoinen ainoastaan perusmuotoistamisen yhteydessä, vaan sitä voidaan käyttää menestyksellä myös taivutusmuotoisten hakemistojen ja kyselyjen sekä karsinnan yhteydessä.

Lopuksi Hollink ym. (2004, 42) tarkastelivat, mikä on yhdyssanojen osittamisen ja karsinnan yhteisvaikutus tuloksiin, kun vertailukohtana käytettiin ilman mitään käsittelyä saatuja tuloksia eli taivutusmuotohakemistolla ennen osittamista ja karsintaa saatuja tuloksia. Tällöin tulokset olivat hollannin 5 prosentista suomen 43 prosenttiin parempia kuin pelkän taivutusmuotohakemiston tulokset. Kaikista neljästä kielestä suomen tuloksellisuudessa tapahtui kaikkein suurin parannus tuon 43 prosentin turvin. Huimasti parantuneesta tuloksellisuudesta huolimatta suomi jäi tuloksellisuudessa kolmeen muuhun kieleen verrattaessa neljännelle ja viimeiselle sijalle. Suomi sijoittui tuloksellisuudessa jo alun alkaen, taivutusmuotohakemistosta haettaessa, huonoimmalle sijalle. (Hollink ym. 2004, 42.)

Nämä tulokset siis tukevat havaintoa, jonka mukaan suomenkielisiä kyselyitä ja hakemistoja on pakko käsitellä jotenkin, jotta saataisiin edes tyydyttäviä tuloksia.

4.4.3 Sananmuotojen käsittelyyn käytettävien menetelmien vertailu toisiinsa

Kun sananmuotojen käsittely osoittautui paremmaksi vaihtoehdoksi kuin käsittelemättä jättäminen, alettiin eri käsittelymenetelmiä ja niillä saatuja tuloksia vertailla ennemminkin toisiinsa kuin taivutusmuodoilla saatuihin tuloksiin. Näin voitiin tutkia eri käsittelymenetelmien keskinäistä paremmuutta. Tässä luvussa esitellään tällaisia vertailuja tehneitä tutkimuksia.

Airion, Keskustalon, Hedlundin ja Pirkolan (2004, 4) tutkimuksessa käytettiin monikielistä kokoelmaa, joka sisälsi 8:lla eri kielellä kirjoitettuja dokumentteja. Monikielisessä testikokoelmassa hakuaiheet esitetään yhdellä lähdekielellä (source language) dokumenttien ollessa usealla kohdekielellä (target-language) kirjoitettuja (Airio, Keskustalo, Hedlund & Pirkola 2003, 1). Tällaisessa monikielisessä (multilingual) dokumenttikokoelmassa on tyypillisesti erilliset hakemistot eri kielille. Näin oli myös tässä tutkimuksessa. Hakemistoja rakennettiin 12 kappaletta 8 hakemiston sijaan, koska englannin, suomen, ruotsin ja saksan kielille laadittiin kaksi erilaista hakemistoa: karsittu ja perusmuotoinen hakemisto. Kyseisten kielten karsitut hakemistot laadittiin Snowball-ohjelmistoilla. Muiden neljän kielen (hollanti, ranska, italia ja espanja) kohdalla käytettiin vain karsittua hakemistoa ja karsinta toteutettiin muilla algoritmeilla. Perusmuoto-ohjelmina käytettiin Lingsoft Oy:n TWOL ohjelmia. Tiedonhakujärjestelmänä oli osittaistäsmäyttävä Inquery. (Airio ym. 2004, 1–4; 9.)

Kielten välisessä tiedonhaussa alkuperäinen eli lähdekielinen kysely käännetään sanakirjan avulla mahdollisimman automaattisesti kohdekieliseksi kyselyksi, jolla etsitään kohdekielellä kirjoitettuja dokumentteja. Airion ym. (2004, 5) tutkimuksessa oli käytössä 10:nen kaksikielistä (bilingual) kyselysarjaa, joiden lisäksi tehtiin kaksi yksikielistä (monolingual) kyselysarjaa. Lähdekielenä oli kaikissa kielipareissa englanti, jolloin alun perin englanninkielisiä kyselyitä käännettiin kullekin kohdekielelle. Yksikielisiä kyselysarjoja ei luonnollisestikaan käännetty. Kaksikielisiä englanti-suomi, englanti-ruotsi ja englanti-saksa kyselysarjoja oli yhteensä kuusi, sillä kyselysarjoista oli sekä karsitut että perusmuotoiset versiot. Perusmuotoistamisen yhteydessä on käytetty myös ositusta, vaikka tämä käy ilmi vasta kyselyjen yksityiskohtaisen tarkastelun yhteydessä. Sen sijaan englanti-hollanti, englanti-ranska, englanti-italia ja englanti-espanja kyselysarjat suoritettiin ainoastaan karsitussa muodossa.

Yksikieliset kyselysarjat olivat englanninkielisiä ja niillä haettiin sekä karsitusta että perusmuotoisesta hakemistosta. (Airio ym. 2004, 1–5.)

Näiden kyselysarjojen saama keskimääräinen tarkkuus vaihteli 17,4 prosentista (englanti-hollanti) 46,3 prosenttiin (yksikielinen englannin karsittu kyselysarja). Kaikista kaksikielisistä kyselysarjoista parhaimman tarkkuuden tuotti perusmuotoinen englanti-suomi kyselysarja, jonka tarkkuus oli 34,0 prosenttia. Kahden englannin yksikielisen kyselysarjan tulokset eivät juuri poikenneet toisistaan, sillä englannin karsitussa hakemistossa suoritettu kyselysarja antoi vain 1,5 prosenttia paremmat tulokset kuin perusmuotoisessa hakemistossa suoritettu kyselysarja. Sen sijaan kaksikielisten englanti-suomi, englanti-saksa ja englanti-ruotsi kyselysarjojen saamat tulokset olivat erilaisia, sillä karsittuja hakemistoja käyttämällä saatiin paljon huonommat tulokset kuin perusmuotoisia hakemistoja käyttämällä. Englanti-ruotsi kyselysarjalla saatiin karsittua hakemistoa käyttämällä 29,9 prosenttia huonommat tulokset kuin perusmuotoisella hakemistolla. Vastaavasti englanti-saksa ja englanti-suomi kyselysarjojen karsitut versiot tuottivat 17,1 ja 44,1 prosenttia huonommat tulokset kuin perusmuotoiset kyselysarjat. (Airio ym. 2004, 5–6.) Karsinnan saamiin huonoihin tuloksiin vaikuttaa kaksikielisiin kyselyihin liittyvät ongelmat kääntämisessä. Osin samat kääntämiseen liittyvät ongelmat ovat läsnä myös perusmuotoisissa kaksikielisissä hauissa, mutta lievempinä, sillä osituksen ansiosta käännetyt kyselyt täsmäytyvät kohdekielisiin dokumentteihin paremmin. Tutkimuksen perusteella karsinnan voidaan sanoa toimineen perusmuotoistamista huonommin. Keskeisin Airion ym. (2004) tulos tämän työn kannalta on se, että perusmuotoistamisen ja karsinnan on todettu olevan tuloksellisuudeltaan lähes yhtä hyvät englanninkielisessä aineistossa yksikielisiä kyselyitä käytettäessä.

Kettunen, Kunttu ja Järvelin (2005, 478) vertailivat suomenkielisessä aineistossa osittaistäsmäyttävää järjestelmää käyttäen neljää eri sananmuotojen käsittelytapaa keskenään: taivutusmuotoja, vartaloita, karsintaa ja perusmuotoja. Heidän tutkimusongelmiaan olivat, millaisia tuloksia perusmuotoisia ja vartalomuotoisia hakuavaimia sisältävät kyselyt saavat toisiinsa nähden ja onko karsinta realistinen vaihtoehto morfologialtaan rikkaiden kielten käsittelyssä. Alatutkimusongelmia olivat muun muassa yhdyssanojen osittaminen ja johdoskyselyt. (Kettunen ym. 2005, 478.) Vartaloita tutkittiin käyttämällä kahta vartalo-ohjelmaa: MaxStemmaa ja Finstemsiä. Lisäksi tutkimuksessa käytettiin kahta erilaista tutkimusympäristöä. MaxStemmaa käytettiin ensimmäisessä ja Finstemsiä toisessa tutkimusympäristössä. (Kettunen ym. 2005, 480–483.) Tässä ei ole kuitenkaan tarpeen käsitellä toista tutkimusympäristöä ja siellä saatuja tuloksia tämän enempää. Vartalomuotoisia hakuavaimia sisältävät kyselyt

muodostettiin katkaisun simulointia käyttämällä, joka kasvattaa kyselyiden avainten määrää todella huomattavasti (Kettunen ym. 2005, 481–483). Katkaisun simuloinnista on kerrottu myös luvussa 4.1.

Testikokoelman dokumenttien relevanssiarviot tehtiin neliportaista relevanssiasteikkoa (3=erittäin relevantti, 2=relevantti, 1=marginaalisesti relevantti ja 0=epärelevantti) käyttäen. Kyseisessä tutkimuksessa näitä edellä mainittuja arvoja yhdisteltiin siten, että dokumenttien relevanssiarviot tehtiin lopulta asteikolla erittäin relevantti (3), relevantti (2–3) ja epärelevantti (1–0). Kun relevanteiksi katsotaan vain arvon 3 saavat dokumentit, puhutaan tiukasta relevanssitasosta. Kun relevanteiksi katsotaan sekä arvon 3 että arvon 2 saaneet dokumentit on kyse niin kutsutusta normaalista relevanssitasosta. (Kettunen ym. 2005, 479.) Niin sanottua liberaalia relevanssitasoa (1–3) ei käytetty lainkaan.

Kettunen ym. (2005, 490) käyttivät Sparck Jonesin (1974, 397) esittelemää menetelmien välisten tilastollisten erojen käytännön vaikutusten arviointitapaa. Jos ero kahden menetelmän välillä on tilastollisesti merkitsevä, voidaan tämän eron käytännön vaikutuksia arvioida seuraavan peukalosäännön avulla:

- Jos menetelmien välinen ero on pienempi kuin 5 prosenttiyksikköä, ei erolla ole käytännössä havaittavia vaikutuksia.
- Jos ero menetelmien välillä on 5-10 prosenttiyksikköä, on eron vaikutus käytännössä havaittava.
- Jos menetelmien välinen ero on yli 10 prosenttiyksikköä, on eron vaikutus käytännössä huomattava. (Sparck Jones 1974, 397.)

Niinpä tulokset esitettiin kertomalla sekä havaitut tilastolliset erot että niiden vaikutukset käytännössä. Perusasetelma tutkimuksen tuloksissa oli se, että perusmuotoistaminen oli paras, vartaloiden tuottaminen toiseksi paras ja karsinta kolmanneksi paras menetelmä. Joskin perusmuotoistaminen oli vain hivenen vartaloiden käyttöä parempi menetelmä. Suurimmat erot perusmuotoja ja vartaloita sisältävien kyselyjen välillä havaittiin, kun perusmuotoisissa kyselyissä oli lisäksi käytetty yhdyssanojen ositusta ja kyselyitä oli laajennettu johdoksilla. (Kettunen ym. 2005, 484–485; 490.)

Kaikki kolme menetelmää saivat paremmat tulokset verrattuna taivutusmuotoiseen hakemiseen. Niinpä perusmuotoistamisella, vartaloilla ja karsinnalla saadut tulokset olivat parempia kuin taivutusmuotoisilla kyselyillä taivutusmuotoisesta hakemistosta saadut tulokset. Erot näillä kolmella menetelmällä saatujen tulosten ja taivutusmuotoisten kyselyiden tulosten välillä olivat tilastollisesti mer-

kitseviä normaalilla ja tiukalla relevanssitasolla. Käytännössä erot olivat kahden parhaiten menestyneen menetelmän ja taivutusmuotoisten kyselyiden välillä huomattavat kummallakin relevanssitasolla. Karsinnan ja taivutusmuotoisten kyselyjen välisen eron ollessa vain havaittava normaalilla ja tiukalla relevanssitasolla tarkasteltuna. (Kettunen ym. 2005, 488–490.)

Entäpä millainen oli karsinnan ja perusmuotoistamisen välinen ero tässä tutkimuksessa? Perusmuotoistamisella saadut tulokset olivat paremmat kuin karsinnalla saadut tulokset. Erot niiden välillä olivat tilastollisesti merkitseviä sekä normaalilla että tiukalla relevanssitasolla. Käytännössä nämä erot olivat havaittavia ainoastaan normaalilla relevanssitasolla, kun taas tiukalla relevanssitasolla erot perusmuotoistamisen ja karsinnan välillä eivät olleet havaittavia. (Kettunen ym. 2005, 488–490.)

Vartaloilla saadut tulokset olivat parempia kuin karsinnalla saadut tulokset ja niiden väliset erot olivat tilastollisesti merkitseviä ja käytännössä havaittavia normaalilla relevanssitasolla. Sen sijaan tiukalla relevanssitasolla karsinnan ja vartaloiden väliset erot eivät olleet tilastollisesti merkitseviä eivätkä havaittavia. (Kettunen ym. 2005, 488–490.)

Saatujen tulosten ohella Kettunen ym. (2005, 491) huomauttivat, että karsinnassa voi olla mahdollista päästä parempiinkin tuloksiin panostamalla suomen kielen käsittelyyn käytettävän karsinta-algoritmin kehittämiseen, sillä nyt käytetty Snowball-ohjelmisto ei ollut suomen käsittelyn kannalta optimaalisin.

5 Tulosjoukkojen päällekkäisyys

Ingwersen (1994) on kirjoittanut polyrepresentaatiosta ja tarkoituksellisen toisteisuuden periaatteesta. Tarkoituksellisen toisteisuuden periaate on yksi polyrepresentaatioteorian takaa löytyvistä periaatteista. Tarkoituksellisen toisteisuuden periaate, jota on sovellettu polyrepresentaatioon, on saanut alkunsa Sparck Jonesin (1990; tässä Ingwersen 1994, 105) esittämistä argumenteista, jotka koskivat toisteisuuden välttämättömyyttä. Toisteisuus on sitä, että samaan käsitteeseen viitataan eri tavoin ja että eri käsitteet yhdistetään käsiteverkon muotoon. Esimerkiksi Turtlen ja Croftin kehittämä päättelyverkkomalli, johon tiedonhakujärjestelmä Inquery perustuu, sallii käsitteeseen viitattavan eri tavoin (Turtle & Croft 1990; tässä Ingwersen 1994, 105).

Tarkoituksellista toisteisuutta on käytetty yksinkertaistetussa muodossa jo vuosikymmeniä operationaalisissa online-ympäristöissä käyttämällä eri indeksointitapoja ja tesauruksia saman viitetietokannan kuvailussa. Kun sama dokumenttikokoelma indeksoidaan eri tavoilla, saadaan kokoelman jokaiselle dokumentille eri dokumenttiversiot. Tätä hienostuneempi ja dynaamisempi tapa toteuttaa toisteisuutta on käyttää käsitteellisiä päättelyverkkoja. Tarkoituksellisen toisteisuuden periaatetta voidaan käyttää erilaisten dokumenttiversioiden luomisessa kuten yllä tai sitä voidaan käyttää luotaessa erilaisia esityksiä käyttäjän kognitiivisesta tilasta. Mitä enemmän informaatio-objektin, esimerkiksi hakupyynnön tai tekstidokumentin, erilaisten versioiden laatimismenetelmät eroavat toisistaan kognitiiviselta alkuperältään, sitä vähemmän eri versioiden saamat tulosjoukot osoittautuvat päällekkäisiksi. Osoituksena siitä voidaan pitää sitä, että löydetyt tulosjoukot olivat melko erilaiset etsittäessä sekä automaattisesti että manuaalisesti indeksoiduista dokumenttikokoelmista. (Ingwersen 1994, 105.) Ingwersenin (1994, 105) mukaan päällekkäisyys jäänee vähäiseksi myös silloin, kun samalla kyselyllä haetaan käyttämällä erilaisia osittaistäsmäyttäviä menetelmiä tai vaihtoehtoisesti erilaisia täystäsmäyttäviä menetelmiä.

Myös Belkin, Cool, Croft ja Callan (1993) ovat sivunneet toisteisuuden periaatetta. Perustelu useiden kyselyversioiden ja tiedonhakumenetelmien käytölle löytyy Belkinin ym. (1993, 339) mukaan tiedonhaussa usein käytetystä probabilistisesta viitekehyksestä, jonka mukaan dokumentin relevanttiuden todennäköisyys kyetään arvioimaan sitä tarkemmin, mitä enemmän kyselyistä ja dokumenteista tai kyselyn ja dokumenttien välisestä suhteesta on saatavilla relevanssista kertovia todisteita. Niinpä jokainen erilainen kyselyversio on uusi todisteiden lähde, jota voidaan käyttää parantamaan ennustetta relevanssin todennäköisyydestä. (Belkin ym. 1993, 339.)

Tulosjoukkojen päällekkäisyyden tutkimusta on motivoinut kaksi asiaa. Tulosjoukkojen päällekkäisyyttä on tutkittu, koska vähäinen päällekkäisyys on toiminut puolestaan yhdistelyn tutkimisen kannustimena. Yhdistelyä onkin tutkittu tiedonhaun piirissä melko runsaasti. Toisekseen on havaittu, että tulosjoukkojen päällekkäisyys vahvistaa relevanttien dokumenttien löytymisen todennäköisyyttä (Pao 1994, 306).

5.1 Päällekkäisyyden ja yhdistelyn läheinen suhde yhdistelyn tutkimisen motivoijana

Päällekkäisyyden tutkiminen liittyy hyvin läheisesti yhdistelyn tutkimiseen. Niitä voidaan tutkia toisistaan erillään, mutta melko tavallisesti tulosjoukkojen vähäinen päällekkäisyys toimii yhdistelyn tutkimisen kannustimena. Päällekkäisyyden ominaisuudet tiedonhaun mittarina selittävät päällekkäisyyden ja yhdistelyn läheisen suhteen. Päällekkäisyys ja uniikkius ovat saman asian kaksi eri puolta. Jos toisiinsa verrattavat tulosjoukot ovat vähäisessä määrin päällekkäisiä, sisältävät ne silloin runsaasti uniikkeja dokumentteja. Tähän perustuu myös yhdistelyn tutkimisen mielekkyys. Mitä vähemmän tulosjoukot sisältävät yhteisiä relevantteja dokumentteja, sitä enemmän ne todennäköisesti sisältävät uniikkeja relevantteja dokumentteja. Lee (1996, 10) arvelee että, mitä vähemmän kahden haun tulosjoukkojen välillä on päällekkäisyyttä, sitä suuremman voidaan olettaa kahden haun yhdistämisestä koituvan tuloksellisuuden parannuksen olevan.

Tässä yhteydessä on myös syytä määritellä, mitä yhdistelyllä tarkoitetaan. Yhdistelyä voidaan tiedonhaun tutkimuksessa tehdä monin eri tavoin, yhdistelemällä erilaisia entiteettejä toisiinsa. Tässä mainitaan kolme keskeistä yhdistelytapaa. **Datafuusiossa** (data fusion) yhdistetään useita tulosjoukkoja, joissa dokumenttien lajitteluarvon laskenta on perustunut samaan dokumenttikokoelmaan. **Kokoelmafuusiossa** (collection fusion) yhdistetään useita tulosjoukkoja, kun kukin tulosjoukko on saatu eri dokumenttikokoelmaa käyttämällä. **Kyselyjen yhdistämisessä** (query combination) yhdistetään useita kyselyesityksiä ennen tiedonhakua. (Braschler 2004, 186.) Äskeisessä yhdistelytapojen jaottelussa yhdisteltävät entiteetit on kuvattu yleisellä tasolla. Kussakin tutkimuksessa yhdisteltävät entiteetit kuvataan tätä yksityiskohtaisemmin, muun muassa datafuusiota tehdessään Lee (1996) on yhdistellyt toisiinsa eri relevanssipalautteilla laajennetuilla kyselyvektoreilla saatuja tulosjoukkoja.

Aiemmissa päällekkäisyyttä tarkastelleissa tutkimuksissa, joita käydään läpi kahdessa seuraavassa alaluvussa, tulosjoukkojen päällekkäisyys on ollut vähäistä. Tämän perusteella tutkimuksissa on päätelty, että on kannattavaa tutkia joko eri kyselyversioilla tai eri dokumenttiversioilla saatujen tulosjoukkojen yhdistelyä. Kolmantena vaihtoehtona on ollut tutkia eri tiedonhakumenetelmillä saatujen tulosjoukkojen yhdistelyä. Yhdistelyä tällä tavoin tarkastelleissa tutkimuksissa on ollut pyrkimyksenä selvittää, saataisiinko eri kysely- tai dokumenttiversioilla tai tiedonhakumenetelmillä saatuja tulosjoukkoja yhdistelemällä tulos, joka ylittää yhdellä ainoalla versiolla tai menetelmällä saadun tulok-

sen. Toiveena on ollut siis tulosjoukkoja yhdistelemällä saada aikaan tiedonhaun tuloksellisuuden kasvamista. (ks. mm. Lee 1996; Belkin ym. 1993.)

5.2 Tulosjoukkojen päällekkäisyys täystäsmäyttävissä järjestelmissä

Tiedonhaun tutkimukset ovat lähestyneet päällekkäisyyttä muun muassa inhimillisen päätöksenteon samankaltaisuuden näkökulmasta sekä tiedonhakujärjestelmän toiminnan samankaltaisuuden näkökulmasta (Saracevic & Kantor 1988, 207). Esimerkiksi ensin mainitusta näkökulmasta sopii Iivosen (1989) tutkimus indeksoijien indeksitermien valinnan johdonmukaisuudesta sekä Saracevicin ja Kantorin (1988) tutkimus hakijoiden eri kyselyversioihin valitsemien hakuavainten päällekkäisyydestä. Esimerkki tiedonhakujärjestelmän toiminnan samankaltaisuuden näkökulmasta on Katzerin, McGillin, Tessierin, Frakesin ja DasGuptan (1982) tutkimus tulosjoukkojen päällekkäisyydestä eri dokumenttiversioista etsittäessä. Päällekkäisyyttä on tutkittu myös tietosisällön näkökulmasta tarkastelemalla, missä määrin eri tietokannat kattavat keskenään samat aihealueet ja miten paljon niiden tietosisällöt menevät päällekkäin (ks. Hood & Wilson 2003). Tämä tutkielma edustaa tiedonhakujärjestelmän toiminnan samankaltaisuuden näkökulmaa, sillä siinä tarkastellaan, missä määrin tulosjoukot osoittautuvat päällekkäisiksi eri kyselyversioita käytettäessä.

Tulosjoukkojen päällekkäisyyden määrää on mahdollista kvantifioida eri tavoin. Muun muassa tässä esiteltävissä aiemmissa tutkimuksissa on käytetty monia erilaisia tapoja. Tavat ovat niin erilaisia, ettei eri tutkimusten päällekkäisyyksiä ole juurikaan mahdollista vertailla keskenään. Yksinkertaisin tapa analysoida tulosjoukkojen päällekkäisyyttä on tehdä pareittaisia vertailuja, jossa kunkin version tulosjoukkoa verrataan vuorotellen muiden versioiden tulosjoukkoihin (Katzer ym. 1982, 265). Lopuksi kunkin pareittaisen vertailun päällekkäisyyksien pohjalta lasketaan keskimääräinen päällekkäisyys (ks. Katzer 1982). Toinen tapa on, että kriteerin täyttävien pareittaisten vertailujen määrää verrataan kaikkien pareittaisten vertailujen määrään. Kriteerinä tätä kvantifiontitapaa käytettäessä on pidetty sitä, että tulosjoukkojen välinen päällekkäisyys on jäänyt alle määritellyn prosenttimäärän (ks. Saracevic & Kantor 1988). Kolmas tapa laskea päällekkäisyyttä on seuraavanlainen. Siinä kaikki tutkimuksen hakuaiheiden etsinnässä löydetyt tulosjoukot ja niiden sisältämät dokumentit muodostavat joukon, jossa tarkastellaan, miten suuri osuus löydetyistä dokumenteista esiintyy kyseisessä dokumenttijoukossa useampaan kertaan (ks. Pao 1994). Neljäs tapa on, että tarkastellaan kaikkia yhden hakuaiheen kaikilla kyselyversioilla saatuja tulosjoukkoja ja katsotaan, kuinka monesta tulosjoukosta

sama dokumentti löytyy (ks. Saracevic & Kantor 1988). Tässä tutkielmassa tehdään tulosjoukkojen pareittaisia vertailuja, jolloin tulosjoukkojen päällekkäisyys lasketaan vertailemalla kullakin kyselyversiolla saatua tulosjoukkoa pareittain muilla kyselyversioilla saatujen tulosjoukkojen kanssa.

Seuraavaksi esitellään aiempia tulosjoukkojen päällekkäisyyttä tarkastelleita tutkimuksia. Ensin esitellään tutkimuksia, joissa päällekkäisyyttä on tutkittu täystäsmäyttäviä järjestelmiä käyttäen ja alaluvussa 5.3 ovat vuorossa tutkimukset, jotka on tehty osittaistäsmäyttäviä järjestelmiä hyödyntäen. Tutkimusten jaottelua niissä käytetyn tiedonhakujärjestelmän perusteella voidaan täydentää jaottelemalla tutkimukset lisäksi sen mukaan, onko niissä ollut keskeisemmällä sijalla hakujen kohdistaminen erilaisiin dokumenttiversioihin vai erilaisten kyselyversioiden käyttö. Erilaisten kyselyversioiden käytön yhteydessä ei dokumenteista laadita useita dokumenttiversioita. Sen sijaan kohdistettaessa hakuja useampiin dokumenttiversioihin joudutaan kulloinenkin kysely sovittamaan vastaavaan dokumenttiversioon sopivaksi eri kyselyversioiden avulla. Siksi puhuttaessa hakujen kohdistamisesta erilaisiin dokumenttiversioihin vilahtaa tekstissä myös eri kyselyversiot.

5.2.1 Tulosjoukkojen päällekkäisyys kohdistettaessa haut erilaisiin dokumenttiversioihin

McGill, Koll & Noreault (1979, 1b-3) tutkivat ensisijaisesti erilaisia lajittelualgoritmeja ja niiden kykyä lajitella Boolen ehdot täyttäneet dokumentit relevanssin todennäköisyyden mukaan laskevaan järjestykseen. He käyttivät tutkimuksessaan osaa ERIC CIJE-viitetietokannasta (Current Index to Journals in Education). Käytetty tietokannan osa sisälsi 10 885 dokumenttia, kun käyttöön oli valittu neljän mediatalon kaikki kahden edellisen vuoden aikana tietokannan ylläpitäjälle toimittamat artikkeliviitteet. (McGill ym. 1979, 61–62.)

Dokumenttiversioksi ei valittu viitetietokannan koko viitetiedostoa, vaan dokumenttiversioiksi valittiin seuraavat osat viitetiedostosta:

1) Nimeke ja kuvaus, joiden sisältönä oli vapaatekstimuotoisia sanoja. Niinpä tätä dokumenttiversiota kutsuttiin vapaatekstiversioksi.
2) Viitteen asiasanakenttä, joka sisälsi kontrolloidusta sanastosta valittuja asiasanoja. Sen vuoksi tätä versiota kutsuttiin joko kontrolloiduksi versioksi tai asiasanaversioksi. (McGill ym. 1979, 61–62.)

Molemmille dokumenttiversioille rakennettiin omat käänteistiedostot. Samassa yhteydessä vapaatekstiversioille tehtiin sulkusanojen poisto. Sen lisäksi vapaatekstiversiot karsittiin karsinta-algoritmilla. Kontrolloidun version asiasanat puolestaan katkaistiin 24. merkin kohdalta. Näin kaukaa tapahtuva katkaisu oli mahdollista, koska asiasanat olivat tässä tutkimuksessa usein fraaseja tai muutoin yhteen kytkeytyneitä sanoja. Tutkimuksessa käytettävät hakuaiheet kerättiin ihmisiltä, joilla oli aito tiedontarve. Keräys toteutettiin pyytämällä käyttäjiä kuvaamaan tiedontarvettaan keräyslomakkeeseen kahden tai kolmen lauseen avulla. Hakupyyntöjä saatiin yhteensä 173 kappaletta. Haut suoritti kolme tiedonhaun ammattilaista, joita ohjeistettiin laatimaan korkeaan saantiin tähtääviä kyselyitä sekä käyttämään kyselyn laadinnassa oikeata sanastoa eli joko vapaatekstisanastoa tai kontrolloitua sanastoa. Ammattilaisten kesken jaettiin sattumanvaraisesti 68 tiedontarvekuvausta siten, että tiedontarvekuvauksen saanut hakija teki haut sekä kontrolloitua että vapaatekstisanastoa käyttäen. Näin ollen sama välittäjä laati kummatkin kyselyversiot. Loput jäljelle jääneet 105 tiedontarvekuvausta jaettiin eri ammattilaisten haettaviksi siten, että toinen ammattilainen käytti hakujen laadinnassa asiasanasanastoa ja toinen vapaatekstisanastoa, joten eri kyselyversiot olivat eri välittäjien laadittavina. (McGill ym. 1979, 62–64; 68–72.)

Tulosjoukkojen päällekkäisyyden tarkasteluun valittiin 33 tiedontarvekuvauksen pohjalta laaditut kyselyt ja niiden tulosjoukot. Vertailussa kummallakin kyselyversiolla saatiin osittain eri dokumentteja sisältävät tulosjoukot, vaikka löydettyjen relevanttien ja epärelevanttien dokumenttien prosenttiosuudet olivat samat kummallakin versiolla. Tutkimuksessa tehtiin havainto, jonka mukaan saman hakijan hakiessa sekä asiasana- että vapaatekstiversiolla, tulosjoukkojen päällekkäisyys oli vain 14 prosenttia. Kun vapaatekstiversio oli eri hakijan haettavana kuin asiasanaversio, oli päällekkäisyys vain 5 prosenttia. Tulosjoukoissa esiintyi toisiinsa nähden siis vain hyvin vähän päällekkäisyyttä. Tutkimuksessa kuitenkin todettiin, ettei tutkimuksessa käytetty aineisto mahdollistanut tämän havainnon syvällisempää tarkastelua, vaan sitä varten tarvittaisiin jatkotutkimusta. (McGill ym. 1979, 75–76.)

Hieman myöhemmin tehdyssä Katzerin, McGillin, Tessierin, Frakesin ja DasGuptan (1982, 262–263) tutkimuksessa vertailtiin keskenään seitsemää dokumenttiversiota. Tarkkuuden ja suhteellisen saannin lisäksi tarkasteltiin myös tulosjoukkojen päällekkäisyyttä (Katzer ym. 1982, 262). Katzerin ym. tutkimus oli kaksivaiheinen, joten tässä esiteltävän ykkösvaiheen perään esitetään lyhyesti myös toisen vaiheen keskeisin sisältö. Aineistona Katzerin ym. (1982, 263) ykkösvaiheen tutkimuksessa käy-

tettiin osaa INSPEC tietokannasta. Tämä osa tietokannasta sisälsi 12 000 artikkeliviitettä, joilla kaikilla oli 7 erilaista dokumenttiversiota, jotka noudattelivat viitetietueen kenttiä. Tiedonhaun ammattilaiset suorittivat kyselyitä tiedontarvitsijoilta kerättyjen 84 tiedontarpeen pohjalta käyttäen täystäsmäyttävää DIATOM-tiedonhakujärjestelmää. Kutakin hakuaihetta etsittiin kohdistamalla kyselyt vuorotellen kuhunkin dokumenttiversioon. Yksi välittäjä laati vain yhteen dokumenttiversioon kohdistettavia kyselyitä. Ammattilaisia kehotettiin laatimaan korkeaan saantiin tähtääviä kyselyitä siten, että haun sai kohdistaa vain yhteen viitetietueen kenttään. Hakujen suorittamisen jälkeen tiedontarvitsijat arvioivat kunkin löydetyn dokumentin relevanttiuden neliportaista arviointia käyttäen: 1=ehdottomasti relevantti, 2=mahdollisesti relevantti, 3=mahdollisesti epärelevantti ja 4=ehdottomasti epärelevantti. (Katzer ym. 1982, 262–263.) Kyselyiden laatimisesta ei kerrota artikkelissa enempää. Dokumenttiversioista, joihin kyselyt kohdistettiin, voidaan kuitenkin päätellä, että kyselyissä on käytetty joko vapaatekstisanoja tai asiasanoja kontrolloidusta sanastosta kulloisestakin dokumenttiversiosta riippuen.

Seuraavassa on listattu kaikki 7 dokumenttiversiota. Kolmessa alimmassa tapauksessa dokumenttiversioksi oli valittu kaksi tietueen kenttää. Kaikista dokumenttiversioista oli poistettu sulkusanat.
TT = nimekkeen vapaateksti sanat
AA = tiivistelmän vapaateksti sanat
DD = kontrolloidusta sanastosta peräisin olevat asiasanat
II = indeksoijan vapaatekstistä valitsemat sanat
Kolme dokumenttiversiota sisältää kaksi kenttää:
TA = nimekkeen ja tiivistelmän vapaateksti sanat eli TT & AA
DI = asiasanat ja indeksoijan valitsemat vapaateksti sanat eli DD & II
ST = nimekkeestä ja tiivistelmästä peräisin olevat vapaatekstisanat karsinta-algoritmilla karsittuina eli karsittu TA (Katzer ym. 1982, 263.)

Tutkimuksen tärkeimmät löydöt olivat, että ensinnäkään erot tuloksellisuudessa eri dokumenttiversioiden välillä eivät olleet suuria. Toisekseen keskimääräisten tulosten valossa tulosjoukkojen väliset päällekkäisyydet olivat pieniä. Myös niiden dokumenttiversioiden tulosjoukkojen välinen päällekkäisyys oli vähäistä, joiden olisi niiden samankaltaisuudesta johtuen olettanut menevän päällekkäin kohtalaisen tai jopa huomattavan paljon. (Katzer ym. 1982, 266; 272.) Katzerin ym. (1982, 266) mukaan yksi pieniä päällekkäisyyksiä selittävä tekijä voisi olla hakijoiden väliset erot, sillä vertailtavista do-

kumenttiversioista toinen oli aina ollut eri ammattilaisen haettavana kuin toinen. Katzer ym. (1982, 266–267) viittaavat kuitenkin McGillin ym. (1979) tekemään tutkimukseen ja siinä esitettyihin tuloksiin, joiden valossa hakijoiden väliset erot eivät näytä olevan ainoa eikä suurin syy sille, että päällekkäisyyttä esiintyy vähän. McGillin ym. (1979) tutkimuksessahan päällekkäisyys oli ollut vähäistä, vaikka haut eri versioilla oli suorittanut sama hakija. (Katzer ym. 1982, 266–267.)

Kolmas Katzerin ym. (1982, 265–266) tekemä löydös oli, että dokumenttien keskimääräinen yhteisesiintyvyys oli suurimmillaan, kun tulosjoukoille yhteisten dokumenttien määrää selvitettiin vertailemalla keskenään tulosjoukkojen erittäin relevantteja osia. Erittäin relevantteista dokumenteista yhteisiksi erittäin relevanteiksi osoittautui nimittäin keskimäärin 35 prosenttia. Sen sijaan yhteisiksi relevanteiksi osoittautui keskimäärin 29 prosenttia toisiinsa verrattavien tulosjoukkojen kaikista relevanteista. Kun yhteisesiintyvyyttä tarkasteltiin kaikkien verrattaviin tulosjoukkoihin sisältyneiden dokumenttien välillä, osoittautui yhteisiksi dokumenteiksi keskimäärin 16 prosenttia. Tulosjoukkojen välinen päällekkäisyys siis väheni, kun tarkasteltavana olevien dokumenttien määrä kasvoi tai tarkemmin sanottuna, kun relevanttius arvioitiin väljemmin. (Katzer ym. 1982, 265–269.) Tämä Katzerin ym. (1982) havainto viittaa siihen, että vaikka eri dokumenttiversioilla saaduissa tulosjoukoissa päällekkäisyyden määrä onkin vähäistä, sisältää juuri tuo päällekkäinen osa suuren määrän relevantteja dokumentteja (Pao 1994, 306).

Katzerin ym. (1982, 262) tutkimuksessa päällekkäisyys- ja tuloksellisuuslukuja käytettiin niiden dokumenttiversioiden tunnistamiseen, joista yhdessä löytyy eniten uniikkeja ja uniikkeja relevantteja dokumentteja. Näin ollen pareittaisten vertailujen lisäksi tehtiin vertailuja, joissa verrattiin toisiinsa yhtä aikaa useamman kuin kahden dokumenttiversion tulosjoukkoja. Silloin tarkasteltiin, paljonko vielä kolmannesta, neljännestä ja viidennestä ja sitä seuraavasta jne. dokumenttiversiosta etsimällä löydettiin dokumentteja, joita aiemmista dokumenttiversioista ei kyetty löytämään. (Katzer ym. 1982, 267–272.) Päällekkäisyys kasvoi johdonmukaisesti aina, kun vertailuun lisättiin uusi dokumenttiversio. Tämä voidaan päätellä seuraavan havainnon perusteella. Viimeiseksi vertailuun mukaan otetusta dokumenttiversiosta löytyneet dokumentit eivät nimittäin pystyneet lisäämään uniikkien dokumenttien määrää enää yhtä paljon kuin tilanteessa, jossa samainen dokumenttiversio otettiin vertailuun ensimmäisenä (Katzer ym. 1982, 269–271). Päällekkäisyyden lisääntymisestä huolimatta jokaisesta dokumenttiversiosta löydettiin myös uniikkeja dokumentteja.

Kun katsottiin, miten se että dokumenttiversioiden tulosjoukkojen päällekkäisyyden vertailuun lisättiin kerralla aina yksi uusi versio, paransi kaikkien relevanttien dokumenttien löytymistä, voitiin noin 70 prosenttia kaikkein relevanteimmista dokumenteista löytää kolmesta dokumenttiversiosta saatuja tulosjoukkoja toisiinsa vertaamalla. (Katzer ym. 1982, 271–272.)

Seuraavaksi esitellään Katzerin & Das-Guptan (1983, 106) kaksivaiheisen tutkimuksen toinen osa. Toisessa vaiheessa käytettiin osaa PsycAbs-viitetietokannasta, joka sisälsi noin 12000 dokumenttia. Hauissa käytettiin täystäsmäyttävää DIATOM-järjestelmää. Edellisestä tutkimusvaiheesta poiketen käytössä oli nyt 4 dokumenttiversiota, 4 kyselyitä suorittavaa tiedonhaun ammattilaista ja 52 kappaletta tiedontarvitsijoilta kerättyjä tiedontarpeita. Käytössä oli neljä erilaista dokumenttiversiota:
DD = Kontrolloidusta sanastosta valitut asiasanat
AA = Tiivistelmän vapaateksti sanat
TT = Nimekkeen vapaateksti sanat
II = Indeksoijan valitsemat vapaateksti fraasit (Das-Gupta & Katzer (1983, 106–107.)

Kukin välittäjä (intermediary) sai kopion hakupyynnöstä ja ohjeet laatia neljä kyselyä kunkin hakupyynnön pohjalta, yhden kuhunkin dokumenttiversioon sopivaksi. Toinen vaihe poikkesi ensimmäisestä siis siten, että sama hakija teki eri dokumenttiversioihin kohdistettavat kyselyt. Löydettyjen dokumenttien relevanssi arvioitiin neliportaisella asteikolla. Muuten tutkimus tehtiin samalla tavoin kuin ensimmäisessäkin tutkimusvaiheessa ja siinä tehdyt keskeiset havainnot olivat samanlaiset kuin aiemmassa tutkimusvaiheessa saadut tulokset. (Das-Gupta & Katzer (1983, 107–108.) Päällekkäisyys vaihteli tällä kertaa 23 prosentista 27 prosenttiin, mikä Das-Guptan ja Katzerin (1983, 107–108) mukaan täsmää aiempien tulosten kanssa. On siis todennäköistä, ettei vähäinen päällekkäisyys selity hakijoiden välisistä eroista käsin, sillä päällekkäisyys oli pientä tälläkin kerralla, vaikka eri dokumenttiversiot olivat saman hakijan haettavana.

Pao (1994, 305–306) analysoi uudelleen Katzerin ym. vuonna 1982 tekemän tutkimuksen dataa sen selvittämiseksi, mikä on tulosjoukoille yhteisten dokumenttien relevantiksi toteamisen todennäköisyys. Todennäköisyys että kahdesta dokumenttiversiosta löydetyt ja tulosjoukoille yhteisiksi osoittautuneet dokumentit arvioitiin joko erittäin relevanteiksi tai relevanteiksi oli suurempi kuin vain yhdestä dokumenttiversiosta löytyneiden dokumenttien todennäköisyys, vertailtiinpa versioiden tulosjoukkoja pareittain miten tahansa (Pao 1994, 309). Kun tulosjoukoille yhteisten dokumenttien relevantiksi ar-

vioimisen todennäköisyyttä arvioitiin tiukalla, normaalilla ja liberaalilla relevanssitasolla, olivat saadut todennäköisyydet tilastollisesti merkitseviä jokaisen pareittaisen vertailun osalta. Todennäköisyys, että tulosjoukoille yhteiset dokumentit arvioitiin relevanteiksi, oli suurin silloin, kun relevanssi arvioitiin tiukalla relevanssitasolla. (Pao 1994, 310–313.)

Pao (1994) kuitenkin huomauttaa, että hänen saamansa tulokset eivät olleet johdonmukaisia eri relevanssitasoilla ja epäilee, että relevanssitasojen suhteen esiintyi luotettavuusongelmia. Sen sijaan että relevantiksi arvioimisen korkeampi todennäköisyys yleistettäisiin koskemaan sekä yhteisiä relevantteja että yhteisiä erittäin relevantteja dokumentteja, on luotettavampaa yleistää se vain yhteisiä erittäin relevantteja dokumentteja koskevaksi. Siksi on luotettavampaa esittää tutkimuksen lopputulemana, että todennäköisyys eri dokumenttiversioille yhteisten dokumenttien arvioimisesta erittäin relevanteiksi on huomattavasti suurempi kuin yhdestä dokumenttiversiosta löytyneiden dokumenttien todennäköisyys. (Pao 1994, 313.) Tästä tulosten yleistämisen varauksellisuudesta huolimatta tulosten suunta on selvä, eikä se vähennä äskeisten tulosten käyttökelpoisuutta tämän tutkielman tarpeisiin.

Myös Pao (1994) esitteli Katzerin ym. tutkimusaineistoon perustuvia päällekkäisyyslukemia, jotka kuitenkin erosivat Katzerin ym. (1982) esittämistä. Tämä selittyy sillä, että Katzer ym. tarkastelivat tulosjoukkojen päällekkäisyyttä ja Pao dokumenttien esiintymisfrekvenssejä. Katzer ym. (1982, 269) esittivät tuloksissa dokumenttiversioilla saatujen tulosjoukkojen pareittaisten vertailujen keskimääräiset epäsymmetriset päällekkäisyydet. Pao (1994, 312) puolestaan laski keskimääräiset esiintymisfrekvenssit ottamalla tarkasteluun kaikki tutkimuksen 84:n hakuaiheen etsinnässä löydetyt tulosjoukot ja niiden sisältämät dokumentit, jolloin voitiin havaita, että 8,85 % kaikista löydetyistä dokumenteista esiintyi kyseisessä dokumenttijoukossa useammin kuin kerran. Vielä harvempi kyseisessä dokumenttijoukossa usein esiintyneestä dokumentista oli relevantti, sillä sellaisten dokumenttien osuus oli 5,3 %. Vastaavasti 3,55 % kyseisessä dokumenttijoukossa useampaan otteeseen esiintyneistä dokumenteista oli erittäin relevantteja dokumentteja. Seuraavaksi tarkasteluun otettiin nämä dokumenttijoukossa monesti esiintyneet dokumentit, jolloin voitiin havaita, että kaikista monesti dokumenttijoukossa esiintyneistä dokumenteista 63 % arvioitiin relevanteiksi ja 42 % erittäin relevanteiksi. Näistä jonkinasteisesti relevanteiksi arvioiduista dokumenteista 67,38 prosenttia osoittautui erittäin relevanteiksi. Nämä löydöt vahvistavat Katzerin ym. (1982) havainnot siitä, että päällekkäisyys on suurinta, kun relevanteiksi katsotaan vain kaikkein relevanteimmat ja pienintä, kun tarkasteluihin otetaan mukaan kaikki dokumentit. (Pao 1994, 312.)

McCain (1989, 110–112) tarkasteli asiasanahauilla ja lähdeteoshauilla saatujen tulosjoukkojen päällekkäisyyttä. Samoilla hakuaiheilla etsittiin erilaisista tietokannoista, sillä asiasanahaut kohdistettiin viitetietokantojen asiasanoihin ja lähdeteoshauilla haettiin viittausindekseistä. McCain (1989, 111) määrittelee lähdeteoshaun siten, että haulla pyritään löytämään ne teokset, joissa siteerataan vanhoja aiheen kannalta relevantteja dokumentteja. Asiasanahakujen ja lähdeteoshakujen tulosjoukkojen keskinäinen vertailu perustuu seuraavanlaiseen oletukseen. Oletuksena on, että artikkelin kirjoittajan valitsemat lähdeteokset ja indeksoijan valitsemat aiheeseen liittyvät asiasanat kattavat hakuaiheen kaksi erilaista puolta (Pao 1984; tässä McCain 1989, 110).

Biolääketieteen tutkijoilta saatuja hakuaiheita oli käytössä yhdeksän kappaletta. Samat tutkijat ehdottivat myös sellaisia aiempia tieteellisiä aikaansaannoksia, joita uudemmat teokset todennäköisesti siteeraavat. Lisäksi he arvioivat hakutulosten relevanssin ja uutuuden (novelty). Asiasanahaut tehtiin MEDLINE, EXCERPTA MEDICA ja PSYCINFO-tietokantoja käyttämällä ja lähdeteoshaut SCISEARCH ja SOCIAL SCISEARCH-tietokantoja käyttämällä. Kaikki haut suoritettiin käyttämällä täystäsmäyttävää DIALOG-tiedonhakujärjestelmää. Kaikkien hakuaiheiden tulosjoukot käytiin läpi siten, että tunnistettiin ne relevantit dokumentit, jotka oli löydetty 1) vain asiasanahaulla, 2) vain lähdeteoshaulla tai 3) molemmilla hakutavoilla. Kun asiasanahakujen ja lähdeteoshakujen tuloksia verrattiin toisiinsa, osoittautui relevanteista dokumenteista molemmille hakutavoille yhteisiksi keskimäärin vain 10 prosenttia. (McCain 1989, 111–113.)

Ingwersen (1994, 105) selitti McCainin (1989) havaitseman vähäisen päällekkäisyyden johtuvan dokumentin kirjoittajan ja indeksoijan erilaisista kognitiivisista rakenteista sekä siitä että kumpikin on tulkinnut tekstejä hyvin erilaisista tavoitteista ja pyrkimyksistä käsin.

5.2.2 Tulosjoukkojen päällekkäisyys käytettäessä erilaisia kyselyversioita

Saracevic ja Kantor (1988, 203) tutkivat saman hakuaihekuvauksen pohjalta laadituista kyselyversioista kahta asiaa: kyselyversioille yhteisten hakuavainten määrää ja eri kyselyversioiden palauttamien tulosjoukkojen päällekkäisyyttä. Tutkimuksessa käytettävänä online-palveluna oli DIALOG, joka oli toiminnoiltaan täystäsmäyttävä. Hakuaiheita oli kaikkiaan 40, joiden pohjalta kyselyitä laati 36 hakijaa. Kun kutakin hakuaihetta kohden laadittiin yhteensä 5 kyselyä, oli kyselyitä yhteensä 200 kappa-

letta. Kukin viidestä kyselyversiosta oli eri hakijan laadittavana. Kyselyjen laatimiseksi kyseisille hakijoille annettiin muistiinkirjoitettu kysymys, joka oli täsmälleen siinä muodossa, jossa tiedontarvitsija oli ongelmansa esittänyt. Hakijoiden käyttöön annettiin myös tesaurus ja muut sopivat välineet, joita he saivat käyttää tarvitessaan. Hakijat saivat myös hioa ja parannella alun perin laatimaansa kyselyä haun tuloksista saamansa palautteen avulla. Näitä viittä kyselyä ja niiden tulosjoukkoja vertailtiin keskenään, mutta ei siis itsensä kanssa, jolloin tehtyjen pareittaisten vertailujen määräksi saatiin 20 vertailua per hakuaihe. Näitä kyselyjen ja tulosjoukkojen pareittaisia vertailuja tehtiin 40:ää hakuaihetta varten yhteensä 800 kappaletta (40x20). Tutkimuksessa esitetyt tulokset on siis saatu laskemalla päällekkäisyyksiä 800 kertaa. (Saracevic & Kantor 1988, 198–203.)

Yleisesti ottaen eri hakijoiden valitsemissa hakuavaimissa oli suhteellisen vähän päällekkäisyyttä. Toisin sanoen, kun hakijoiden piti laatia kyselyt saman hakuaiheen pohjalta, heidän tekemänsä valinnat osuivat samoihin hakuavaimiin vain muutaman hakuavaimen kohdalla suurimman osan hakuavaimista ollessa erilaisia muiden hakijoiden valitsemien avainten kanssa. Kyselyjen pareittaisia vertailuja, joissa hakuavainten päällekkäisyys jäi alle 25 prosentin, oli kaikista pareittaisista vertailuista yli puolet (56,4 %). Keskimäärin hakuavainvalinnat menivät päällekkäin 27 prosenttia. (Saracevic & Kantor 1988, 203–204.) Tämä vahvistaa entisestään päätelmää, jonka mukaan eri hakijat näkevät samassa hakuaihekuvauksessa eri asioita, jolloin he myös tulkitsevat sitä eri tavalla. Tämä taas näkyy kyselyiden laadinnassa siten, että hakijat valitsevat kyselyihin toisistaan poikkeavia kielellisiä ja loogisia rakenteita, jolloin kyselyt löytävät tietokannasta eri dokumentteja. (Saracevic & Kantor 1988, 204.)

Tulosjoukkojen päällekkäisyyden tarkastelussa verrattiin, kuinka paljon samoja dokumentteja kaksi tulosjoukkoa sisälsi. Tarkasteluja tehtiin kahdella tavalla. Ensimmäisessä tarkastelussa yhteisten dokumenttien määrää tarkasteltiin kokonaisissa tulosjoukoissa. Toisessa tarkastelussa katsottiin, missä määrin tulosjoukkojen relevantit ja osittain relevantit dokumentit osoittautuivat kummallekin tulosjoukolle yhteiseksi, kun tarkastelun ulkopuolelle rajattiin tulosjoukon epärelevantit dokumentit. Ensin esitetään tulokset, joissa yhteisten dokumenttien määrää tarkasteltiin kokonaisissa tulosjoukoissa. Pareittaisia vertailuja, joissa tulosjoukkojen dokumenteista yhteisiksi osoittautui alle 5 prosenttia, oli kaikista pareittaisista vertailuista yli puolet eli 58,6 prosenttia. Pareittaisia vertailuja, joissa yhteisten dokumenttien osuus oli enemmän kuin 5 prosenttia mutta alle 25 prosenttia, oli kaikista pareittaisista vertailuista 20 prosenttia. Vertailtaessa tulosjoukkoja pareittain relevanttien ja osittain relevanttien

osalta oli yhteisten dokumenttien osuus alle 5 prosenttia yli puolessa pareittaisista vertailuista (58,9 %). Päällekkäisyyden määrä oli enemmän kuin 5 prosenttia mutta vähemmän kuin 25 prosenttia 17 prosentissa pareittaisista vertailuista. Keskimäärin löydetyistä dokumenteista osoittautui samoiksi 17 prosenttia, kun päällekkäisyyttä tarkasteltiin kokonaisissa tulosjoukoissa ja 18 prosenttia kun asiaa tarkasteltiin relevanttien ja osittain relevanttien dokumenttien osalta. Niinpä päällekkäisyyttä esiintyi tulosjoukkojen välillä huomattavasti vähemmän kuin hakuavainten valinnassa. (Saracevic & Kantor 1988, 204–205.)

Koska päällekkäisyys hakuavainten valinnassa oli suurempaa kuin päällekkäisyys verrattavissa tulosjoukoissa, tutkimuksessa selvitettiin, selittääkö hakuavainten suuri päällekkäisyys dokumenttien yhteisesiintymisen tulosjoukoissa. Vastauksena oli, ettei näin ole, sillä vain 2,5 prosenttia löydettyjen dokumenttien yhteisesiintymisestä voitiin johtaa hakuavainten päällekkäisyydestä. (Saracevic & Kantor 1988, 204.) Koska tulosjoukkojen päällekkäisyys oli yhteensä noin parisenkymmentä prosenttia, selitti hakuavainvalintojen päällekkäisyys siitä vain pienen osan. Yleisesti ottaen voidaan siis sanoa, ettei hakuavainten ja tulosjoukkojen päällekkäisyys liity läheisesti toisiinsa täystäsmäytyksessä (Saracevic & Kantor 1988, 204).

Saracevicin ja Kantorin (1988, 205) tutkimuksessa tarkasteltiin lisäksi sitä, monestako saman hakuaiheen pohjalta laaditun kyselyn tulosjoukosta dokumentit löytyvät. Noin 79 prosenttia kaikista löydetyistä dokumenteista, olivatpa ne sitten relevantteja, osittain relevantteja tai epärelevantteja, löydettiin ainoastaan yhdellä hakuaiheen pohjalta laaditulla kyselyllä (ks. Taulukko 3). Kahdella saman hakuaiheen pohjalta laaditulla kyselyllä niistä löydettiin 13 prosenttia, kolmella kyselyllä löydettiin 6 prosenttia ja neljällä tai useammalla kyselyllä löydettiin 2 prosenttia. Relevanttien ja osittain relevanttien osalta vastaavat lukemat olivat sellaiset, että 72 prosenttia löydetyistä relevanteista ja osittain relevanteista löydettiin yhdestä tulosjoukosta, 17 prosenttia kahdesta tulosjoukosta, 8 prosenttia kolmesta tulosjoukosta ja 3 prosenttia neljästä tai useammasta tulosjoukosta. Kaikista löydetyistä epärelevanteista dokumenteista 86 prosenttia löydettiin vain yhdestä tulosjoukosta, 10 prosenttia kahdesta, 3 prosenttia kolmesta ja 1 prosentti neljästä tai useammasta tulosjoukosta. (Saracevic & Kantor 1988, 205–207.) Taulukon 3 perusteella voidaan sanoa, että Saracevicin ja Kantorin tulosten mukaan sekä tulosjoukkojen relevantit että epärelevantit osoittautuivat vähäisessä määrin verrattaville tulosjoukoille yhteisiksi.

TAULUKKO 3. Dokumenttien esiintyminen yhden hakuaiheen kyselyversioilla saaduissa tulosjoukoissa.

	1 tulosjoukosta	2 tulosjoukosta	3 tulosjoukosta	4 tai useammasta tulosjoukosta	Yhteensä
Kaikki löydetyt	79 %	13 %	6 %	2 %	100 %
Relevantit ja osittain relevantit	72 %	17 %	8 %	3 %	100 %
Epärelevantit	86 %	10 %	3 %	1 %	100 %

Äskeiseen tarkasteluun pohjautuen Saracevic ja Kantor (1988) totesivat, että mitä useammalla kyselyllä dokumentti löydettiin, sitä todennäköisemmin dokumentti oli relevantti. Mitä useammin saman hakuaiheen pohjalta laaditut kyselyt palauttivat tuloksena saman dokumentin, sitä todennäköisemmin se oli relevantti. Tätä havaintoa Saracevic ja Kantor pitivät yhtenä tutkimuksensa tärkeimpänä löytönä. (Saracevic & Kantor 1988, 205–207.)

5.3 Tulosjoukkojen päällekkäisyys osittaistäsmäyttävissä järjestelmissä

Seuraavaksi esiteltävä tutkimus poikkeaa yllä esitellyistä siinä käytetyn tutkimusaineiston perusteella. Leen (1996, 2–7) tutkimuksen tutkimusaineistona käytettiin TREC-testikokoelmaa ja vektorimalliin perustuvaa SMART-tiedonhakujärjestelmää. Hänen käyttämänsä TREC-testikokoelma sisälsi valmiiden hakuaiheiden lisäksi yli 740 000 kokotekstiartikkelia (Lee 1996, 7).

Lee (1996) tutki sekä päällekkäisyyttä että yhdistelyä, joista tässä keskitytään enimmäkseen päällekkäisyyttä tarkastelleeseen osioon. Tutkimuksessa erilaiset kyselyversiot laadittiin kyselyvektoreita eri relevanssipalautteilla laajentaen. Erilaisia relevanssipalautteita oli käytössä viisi kappaletta. Eri relevanssipalautteilla laajennettujen kyselyvektoreiden samankaltaisuutta tarkasteltiin kahdella eri tavalla. (Lee 1996, 2–7.) Ensin tarkasteltiin alkuperäisten ja laajennettujen kyselyvektorien samankaltaisuutta toisiinsa nähden. Eri relevanssipalautteilla laajennetut kyselyvektorit olivat melko erilaisia alkuperäisiin kyselyvektoreihin verrattuna. Toisekseen tutkittiin laajennettujen kyselyvektorien välistä samankaltaisuutta, joka osoittautui melko suureksi. Tämän jälkeen katsottiin, missä määrin erilaisilla laajennetuilla kyselyvektoreilla saadut tulosjoukot sisälsivät samoja dokumentteja. Tämä tehtiin vertailemalla tulosjoukkoja pareittain. Oletuksena oli, että mitä erilaisemmat laajennetut kyselyvekto-

rit ovat, sitä vähemmän niillä saatujen tulosjoukkojen pitäisi omata yhteisiä dokumentteja. (Lee 1996, 7–9.)

Lee (1996, 8) raportoi erilaisilla relevanssipalautteilla laajennetuilla kyselyvektoreilla löydetyn melko erilaiset tulosjoukot, vaikka kaikki eri kyselyvektorit olivat tuloksellisuudeltaan samantasoisia. Tämä tutkimuksessa esitetty toteamus on hämmästyttävä siinä mielessä, että tulosjoukkojen päällekkäisyys vaihteli kymmenessä pareittaisessa vertailussa 52 prosentista jopa 99 prosenttiin, kun päällekkäisyyttä tarkasteltiin kokonaisissa tulosjoukoissa, joten tutkimuksessa saatuja tulosjoukkoja pitäisi ennemminkin luonnehtia melko tai hyvin samankaltaisiksi. Noista kymmenestä pareittaisesta vertailusta kolmessa päällekkäisyys oli 52–54 prosenttia. Kolmen muun pareittaisen vertailun osalta päällekkäisyys oli 64–67 prosenttia. Yhdessä pareittaisessa vertailussa päällekkäisyys oli 74 %. Kahdessa pareittaisessa vertailussa päällekkäisyys oli 88 %. Yhden pareittaisen vertailun päällekkäisyys oli 99 %. Eniten päällekkäisyyttä esiintyi tosiaankin verrattaessa toisiinsa tulosjoukkoja, jotka oli saatu vektoreilla, jotka olivat keskenään samankaltaisimmat. (Lee 1996, 7–9.)

Lisäksi Lee (1996, 1–11) tarkasteli, millainen vaikutus laajennettujen kyselyvektorien tulosjoukkojen yhdistämisellä on tiedonhaun tuloksellisuuteen. Melko suuresta tulosjoukkojen välisestä päällekkäisyydestä huolimatta Lee (1996, 10–11) onnistui saamaan laajennettujen kyselyvektoreiden tulosjoukkojen yhdistelyllä paremmat tulokset kuin laajennetuilla kyselyvektoreilla yksinään.

5.4 Päällekkäisyys ja dokumenttien positiot tulosjoukoissa

Harmanin (1987) tutkimus toimii esimerkkinä tulosjoukon dokumenttien lajittelusta, dokumenttien positioista tulosjoukossa ja niissä sananmuotojen käsittelyn seurauksena tapahtuvista muutoksista. Harman (1987) vertaili karsinnan aiheuttamia position muutoksia taivutusmuotoisilla kyselyillä löydettyjen dokumenttien positioihin. Tässä esitettävä esimerkki positioista ja niissä tapahtuvista muutoksista on olennainen siksi, että päällekkäisyyden määrän tarkastelu toteutetaan tässä tutkielmassa konkreettisesti juuri dokumenttien positioiden avulla. Tässä tutkimuksessa on tarkoitus muun muassa vertailla karsinnalla löydettyjen dokumenttien positioita perusmuotoistamisella löydettyjen dokumenttien positioihin ja katsoa päätyvätkö samat dokumentit positioihin, jotka sisältyvät saman katkaisupisteen rajaamaan tulosjoukkoon (ks. tarkemmin luku 6.4.1).

Harman (1987, 105) otti yksityiskohtaiseen analyysiin useita kyselyitä sen selvittämiseksi, miten karsinta vaikuttaa dokumenttien sijoitukseen tulosjoukossa. Yksityiskohtaiseen tarkasteluun otettiin muun muassa englanninkielisen Cranfield-kokoelman kysely 16. Kyselyllä 16 oli dokumenttikokoelmassa 4 relevanttia dokumenttia, joiden tunnistenumerot olivat 498, 106, 266 ja 196. Sen enempää kyseisestä kyselystä ei kerrotakaan. Cranfield-kokoelman kyselyistä kerrotaan kuitenkin, että taivutusmuotoiset kyselyt sisälsivät keskimäärin 10 hakuavainta. Lisäksi tiedetään, että Porter-algoritmilla karsittujen kyselyiden karsintavartalomuotoiset hakuavaimet kattoivat keskimäärin 26 sananmuotoa. Tutkimuksessa Porter-algoritmin lisäksi käytetyllä S-algoritmilla karsitut hakuavaimet kattoivat puolestaan keskimäärin 15 sananmuotoa. Kun kysely 16 suoritettiin taivutusmuotoisena, olivat kyseisten dokumenttien sijoitukset tulosjoukossa 1 (498), 2 (106), 29 (266) ja 36 (196). (Harman 1987, 105–106.)

Ensimmäinen näistä dokumenteista, joka oli tunnistenumeroltaan 498, sisälsi 17 kyselyyn täsmäävää avainta, joista uniikkeja avaimia oli seitsemän. Lisäksi dokumentin neljällä avaimella oli dokumentissa taipuneita muotoja, mikä nostaisi kyseisen dokumentin sijoitusta tulosjoukossa karsintaa käytettäessä, mikäli dokumentti ei jo olisi tulosjoukon ensimmäisenä. Toisen relevantin dokumentin, dokumentin 106, sisältämillä avaimilla ei ollut dokumentissa yhtäkään taipunutta muotoa, joten sen sijoitus tulosjoukossa ei oletettavasti tulisi muuttumaan karsintaa käytettäessä. Tämä dokumentti kuitenkin, sen lisäksi, että se oli lyhyt, sisälsi 10 kyselyyn täsmäävää hakuavainta, joista uniikkeja avaimia oli 5, joten se sijoittui jo valmiiksi tulosjoukossa korkealle. Kolmas dokumentti, tunnistenumeroltaan dokumentti 266, sisälsi 8 kyselyyn täsmäävää hakuavainta, joista kolme oli uniikkeja. Lisäksi dokumentissa oli 8 taivutusmuotoa näistä kyselyyn täsmäävistä hakuavaimista. Sen sijoitus tulisi nousemaan karsintaa käytettäessä huomattavasti. Neljäs dokumentti (196) sisälsi myös 8 kyselyyn täsmäävää hakuavainta, joista uniikkeja avaimia oli 4. Lisäksi dokumentti sisälsi yhden taipuneen avaimen, jolloin sen sijoitus nousisi hieman karsintaa käytettäessä. (Harman 1987, 105.)

Relevantin dokumentin lopullinen sijoitus tulosjoukossa ei määräydy ainoastaan sen oman, lajittelualgoritmin sille antaman painoarvon perusteella, vaan myös epärelevanttien dokumenttien sijoitukset tulosjoukossa vaikuttavat sen lopulliseen sijoitukseen. Sekä relevanttien että epärelevanttien dokumenttien sijoitukset tulosjoukossa olivat erilaiset eri karsinta-algoritmeja käytettäessä, sillä eri karsinta-algoritmeilla karsitut hakuavaimet kattoivat eri määriä sananmuotoja. Ensin käydään läpi Porter-algoritmin vaikutus dokumenttien positioihin. Porter-algoritmilla saadut karsintavartalot kattoivat

paljon enemmän sananmuotoja kuin S-algoritmi, jolloin useiden epärelevanttien dokumenttien sijoitukset nousivat niiden taivutusmuotoisilla hauilla saamiin sijoituksiin verrattuna. Niinpä Porter-algoritmia käytettäessä dokumentti 106 putosi sijalta 2 sijalle 7 ja dokumentti 196 putosi sijalta 36 sijalle 79. Sen sijaan dokumentin 266 sijoitus nousi sijalta 29 sijalle 9 sen ansiosta, että Porterin algoritmilla karsitut hakuavaimet onnistuivat kattamaan kaksi hyödyllistä sananmuotoa, joita taivutusmuotoinen kysely ei ollut kattanut. Dokumentti 498 oli tulosjoukossa sijalla 1, joten sen sijoitus pysyi muuttumattomana. Porter-algoritmiin verrattuna S-algoritmilla käsitellyt hakuavaimet kattoivat vähemmän sananmuotoja ja samalla myös nostivat vähemmän epärelevanttien dokumenttien positioita antaen relevanteille dokumenteille positiot 1, 2, 26 ja 39. (Harman 1987, 105.)

5.5 Päällekkäisyyden tutkimisen tarve

Yleensä tiedonhaun tuloksia mitataan ja esitetään saannin ja tarkkuuden avulla. Saadut saanti- ja tarkkuusarvot esitetään usein piirtämällä niiden pohjalta saanti-tarkkuuskäyrät. Kaksi tiedonhakumenetelmää voi kuitenkin saada identtiset saanti-tarkkuuskäyrät, mutta silti palauttaa tulosjoukot, jotka sisältävät täysin eri dokumentit. Tavanomaiset tuloksellisuusmittarit (saanti ja tarkkuus) eivät vain kykene tuomaan tällaista tietoa esille. (McGill ym. 1979, 13.) Niinpä kaksi erilaista menetelmää voidaan katsoa samankaltaisiksi vasta, kun ne ovat tuloksellisuudeltaan samanlaiset ja ne noutavat samat dokumentit (Das-Gupta & Katzer 1983, 106). Tämä pätee myös perusmuotoistamiseen ja karsintaan, joiden tiedetään neljännen luvun perusteella olevan tuloksellisuudeltaan melko samanlaiset. Tässä tutkimuksessa selvitettäväksi jää se, noutavatko perusmuoto-ohjelmalla ja karsinta-algoritmilla käsitellyt kyselyt perusmuotoisista ja karsituista hakemistoista keskenään samat vai eri dokumentit. Silloin nähdään, ovatko ne menetelminä keskenään samankaltaisia vai erilaisia.

Nyt tehtävä tutkimus on tutkimusasetelmaltaan samankaltainen kuin osittaistäsmäyttävällä järjestelmällä päällekkäisyyttä luvussa 5.3 tarkastelleella tutkimuksella. Tästä huolimatta aiempien tutkimusten kyselyversioiden ja tässä käytettävien kyselyversioiden erot tekevät tästä tutkielmasta erilaisen kaikkiin yllä referoituihin aikaisempiin tutkimuksiin nähden. Tätä eroa kuvataan lähemmin luvussa 5.6. Lisäksi erona on se, että aiemmat tutkimukset ovat tutkineet päällekkäisyyttä englanninkielistä aineistoa käyttäen. Tämä tutkimus on tietääkseni ainoa, jossa päällekkäisyyden määrää tutkitaan myös suomenkielisessä aineistossa.

5.6 Tutkimusongelma

Yllä esitellyissä täystäsmäyttäviä järjestelmiä käyttäneissä tutkimuksissa hakupyynnön sisältämän käsitteen pohjalta on valittu erilaisia ilmauksia, joita on käytetty hakuavaimina kyseisen hakupyynnön pohjalta laadituissa kyselyversioissa. Samaan käsitteeseen viittaavien ilmaisujen väliset suhteet ovat voineet olla hierarkkisia, synonyymisia ja assosiaatioon perustuvia. Tämä on mahdollistanut, etteivät eri versioiden hakuavain valinnat ja sen myötä tulosjoukot ole juurikaan olleet päällekkäisiä. Näin on ollut myös niissä tutkimuksissa, joissa on ollut keskeisemmällä sijalla hakujen kohdistaminen johonkin dokumenttiversioon.

Yllä referoidussa Leen (1996) osittaistäsmäyttävää järjestelmää käyttäneessä tutkimuksessa erilaiset kyselyversiot laadittiin kyselyvektoreita eri relevanssipalautteilla laajentaen. Relevanssipalautetta käytettäessä tulosjoukon kärkipään dokumenttien oletetaan olevan relevantteja, jolloin kyselyjä laajennetaan näiden dokumenttien sanoilla. Järjestelmä siis poimii hakijan hyviksi dokumenteiksi arvioimista dokumenteista uusia hakuavaimia ja muokkaa kyselyä niiden avulla uusiksi. Relevanssipalautteen hyödyntämiseen sisältyy tyypillisesti myös huonojen hakuavainten poistaminen. Tämän vuoksi relevanssipalautteella laajennetut kyselyvektorit poikkeavat alkuperäisistä laajentamattomista kyselyvektoreista enemmän kuin karsitut, osittamattomat perusmuotoiset ja ositetut perusmuotoiset kyselyt poikkeavat alkuperäisistä käsittelemättömistä kyselyistä.

Tämän opinnäytetyön kannalta on olennaista selvittää, onko tulosjoukkojen päällekkäisyys vähäistä myös silloin, kun eri kyselyversioissa on käytössä periaatteessa samat ilmaisut, vaikka näiden ilmaisujen muodot poikkeavat toisistaan perusmuoto-ohjelmalla ja karsinta-algoritmilla tehdyn käsittelyn vuoksi. Toisin sanoen on tarpeen tarkastella, missä määrin päällekkäisyyttä esiintyy, kun käytetyt hakuavaimet ovat samat, mutta poikkeavat toisistaan muotonsa ja sisältämiensä merkkien määrän perusteella. Kyselyversioiden hakuavainten lähes täydellisen samankaltaisuuden vuoksi voidaan olettaa, ettei päällekkäisyyden määrä jää tässä tutkimuksessa yhtä pieneksi kuin aiemmissa tutkimuksissa, joissa hakuavainten päällekkäisyys on ollut pienempää.

Tässä tutkielmassa ei verrata perusmuoto-ohjelmilla ja karsinta-algoritmeilla käsiteltyjä kyselyitä alkuperäisiin kyselyihin niiden välisen samankaltaisuuden selvittämiseksi. Tässä ei myöskään verrata eri kyselyversiota toisiinsa sen selvittämiseksi, missä määrin kyselyt ovat keskenään samanlaisia niil-

le tehdystä käsittelystä huolimatta. Tässä tutkielmassa lähdetään liikkeelle siitä, että kyselyjen välinen samanlaisuus on niin suurta, ettei sitä kannata edes tutkia. Sen sijaan tarkoituksena on tutkia, miten paljon päällekkäisyyttä esiintyy, kun karsituilla kyselyillä tietokannan karsitusta hakemistosta saatuja tulosjoukkoja ja perusmuotoisilla kyselyillä perusmuotoisesta hakemistosta saatuja tulosjoukkoja verrataan keskenään. Tätä tutkitaan sekä suomen- että englanninkielistä aineistoa käyttäen. Lisäksi suomenkielisessä aineistossa on tarkoituksena tutkia, missä määrin tulosjoukot ovat päällekkäisiä, kun verrataan keskenään ositettujen ja osittamattomien kyselyversioiden tulosjoukkoja.

Päällekkäisyyden määrän tarkastelu toteutetaan konkreettisesti dokumenttien positioiden avulla. Sen vuoksi alla esitetään seuraavanlaiset oletukset siitä, miten kyselyjen karsinnan, perusmuotoistamisen tai osituksen oletetaan vaikuttavan dokumenttien positioihin. Luvun 5.4 Harman (1987) esimerkissä dokumenttien positiot tulosjoukoissa olivat erilaiset, kun verrattiin karsituilla kyselyillä saatujen dokumenttien positioita taivutusmuotoisilla kyselyillä saatujen dokumenttien positioihin. Vertailu tapahtuu tässä tutkielmassa perusmuotoisten, ositettujen perusmuotoisten ja karsittujen kyselyjen dokumenttien positioiden välillä. Ensimmäiseksi esitetään oletuksia, jotka liittyvät perusmuotoisilla ja karsituilla kyselyillä löytyvien dokumenttien positioihin ja lajitteluun. Suomenkielisessä aineistossa perusmuotoisten hakuavainten ei oleteta muuttavan dokumenttien sijoituksia yhtä voimakkaasti kuin karsittujen hakuavainten, koska ne ovat merkkijonoina karsittuja hakuavaimia pidempiä, jolloin ne oletettavasti täsmäytyvät harvempiin avaimiin. Tämän suhteellisen maltillisen käyttäytymisen ansiosta, perusmuotoisten avainten ei oleteta nostavan myöskään epärelevanttien dokumenttien sijoituksia yhtä voimakkaasti kuin karsittujen hakuavainten.

Ositetut perusmuotoiset hakuavaimet ovat pituudeltaan kaikkein lyhyimpiä, mutta niitä ei käytetä kyselyissä yksinään, koska vain yhdyssanat ositetaan. Lisäksi kyselyissä käytettävä yhdyssanaoperaattori estää, ettei ositettu perusmuotoinen hakuavain voi täsmäytyä muihin kuin yhdyssanoihin. Yhdyssanaoperaattorin vaikutuksesta ositettu perusmuotoinen hakuavain täsmäytyy yhdyssanoihin, jossa haetut yhdysosat sijaitsevat samassa sanassa annetussa järjestyksessä, mutta hakuavaimen kanssa täsmäävän yhdyssanan alussa, välissä tai lopussa voi olla muitakin yhdysosia. Esimerkiksi hakuavain #0(pankki kortti) täsmäytyy myös sanan pikapankkikortti kanssa. Tämän pohjalta voidaan esittää oletuksia ositetuilla perusmuotoisilla hakuavaimilla saatavien dokumenttien positioista. Niinpä seuraavaksi verrataan ositettujen perusmuotoisten hakuavainten ja kahden muun hakuavaintyypin tulosjoukon lajitteluun liittyviä oletuksia toisiinsa. Suomenkielisessä aineistossa ositettujen perusmuotois-

ten yhdyssanojen oletetaan täsmäytyvän useampiin sanoihin kuin perusmuotoisten ja karsittujen yhdyssanojen. Esimerkiksi karsittu yhdyssana täsmäytyy vain samanalkuisiin yhdyssanoihin. Tämän seurauksena ositetut perusmuotoiset hakuavaimet saattavat täsmäytyä relevantteihin ja epärelevantteihin dokumentteihin perusmuotoisia ja karsittuja avaimia paremmin ja nostaa dokumenttien sijoituksia tulosjoukossa enemmän kuin perusmuotoiset ja karsitut hakuavaimet.

Sekä karsitut yhdyssanat että ositetut perusmuotoiset yhdyssanat saattavat tuoda kyselyyn mukaan yhden aiheen lisäaspektin, mikä voi olla tiedonhaun kannalta joko hyvä tai huono asia. Kumpikin vaihtoehto on mahdollinen etenkin silloin, kun hakuavain täsmäytyy sanaan, jonka perässä on yhdysosa, jota ei eksplisiittisesti etsitty. Näin on esimerkiksi silloin kun, ositettu perusmuotoinen hakuavain #0(asunto kauppa) tai karsittu hakuavain asuntokaup täsmäytyy yhdyssanan asuntokaupparii-ta kanssa. Tämän perusteella voidaan olettaa, että karsitut yhdyssanat täsmäytyvät relevantteihin ja epärelevantteihin dokumentteihin osittamattomia perusmuotoisia yhdyssanoja paremmin.

Viitteitä siitä, että englanninkielisessä aineistossa perusmuotoiset ja karsitut hakuavaimet toimivat hyvin samankaltaisesti, antavat hyvin pienet erot perusmuotoisten ja karsittujen hakujen tuloksellisuudessa. Siksi on vaikea sanoa etukäteen kummat hakuavaimet oletettavasti muuttavat dokumenttien sijoituksia enemmän käytettäessä testikokoelmana englanninkielistä TRECiä.

6 Tutkimusaineisto ja menetelmät

6.1 Tiedonhakujärjestelmä

Tutkimuksessa tiedonhakujärjestelmänä käytetään Massachussetin yliopistossa kehitettyä osittaistäsmäyttävää Inquery-tiedonhakujärjestelmää, joka perustuu probabilistiseen päättelyverkkoon. Inquery on suunniteltu suurten tietokantojen käyttämiseen, joten se soveltuu hyvin myös tässä tutkimuksessa käytettäväksi (Callan, Croft & Harding 1992, 1). Inquerysta on tutkimuksessa käytössä versio 3.1. Käytettävä tiedonhakujärjestelmä päättelee dokumenttien ja kyselyiden sisällön avulla dokumenttien ja kyselyiden välisiä suhteita sekä arvioi sen todennäköisyyttä, että käyttäjän kyselynä tai kyselyinä ilmaistu tiedontarve täyttyy käytettäessä dokumenttia evidenssinä (Turtle & Croft 1990, 1). Tietokan-

nan jokainen dokumentti sisältää useita vihjeitä relevanssista. Päättelyverkoissa kyselyn ja dokumentin sisällöstä kertovia useita evidenssin lähteitä yhdistellään relevanssin todennäköisyyden arvioimiseksi. (Rajashekar & Croft 1995, 272.)

Inquery käyttää merkkijonojen painotukseen tf.idf kaavaa tai oikeammin sen muunnosta, sillä Inqueryssa merkkijonojen laskemiseen käytetään termifrekvenssin, dokumenttifrekvenssin ja käänteisen dokumenttifrekvenssin lisäksi tietoja dokumentin pituudesta, tietokannan dokumenttien keskimääräisestä pituudesta sekä tietokannan dokumenttien kokonaismäärästä. Nämä tiedot normalisoivat avainten saamia painoja. Termifrekvenssi tarkoittaa avaimen esiintymisfrekvenssiä dokumentissa, dokumenttifrekvenssi tarkoittaa avaimen sisältävien dokumenttien määrää tietokannassa ja käänteinen dokumenttifrekvenssi saadaan, kun tietokannan koko jaetaan avaimen sisältävien dokumenttien määrällä. Tf.idf painotus painottaa avainta, joka on yleinen dokumentissa, mutta harvinainen tietokannassa.

Inqueryn versiossa V3.1 lasketaan avaimen t paino käyttämällä seuraavaa kaavaa:

$$0{,}4 + 0{,}6 * \left(\frac{tf_{t,d}}{tf_{t,d} + 0{,}5 + 1{,}5 * \frac{length(d)}{avglen}} \right) * \left(\frac{\log \frac{N + 0{,}5}{n_t}}{\log N + 1} \right)$$

$tf_{t,d}$ = avaimen t esiintymisfrekvenssi dokumentissa d

length(d) = dokumentin d pituus sanoissa mitattuna

avglen = tietokannan dokumenttien keskimääräinen pituus sanoissa mitattuna

N = dokumenttien määrä tietokannassa

n_t = dokumenttien määrä, jotka sisältävät avaimen t

0.4 ja 0.6 ovat vakioita, jotka normalisoivat tf*idf-painotusta.

(Allan ym. 1997, 170.)

Tällä kaavalla avaimet saavat painot, jotka ovat väliltä 0–1.

Inquerylle syötettävät kyselyt voidaan laatia joko luonnollisen kielen avulla tai rakenteista kyselykieltä käyttämällä. Luonnollisen kielen kyselyt muutetaan Inqueryn syntaksiin sopiviksi #sum-

operaattorin avulla. (Callan ym. 1992, 6.) Inqueryn kyselykielen operaattorit ovat Boolen loogisten operaattoreiden kaltaisia. Kolmen yleisimmän operaattorin, #and-, #or- ja #not-operaattorin, lisäksi Inqueryn kyselykielessä on monia muitakin operaattoreita, joista tässä käydään läpi yleisimmin kyselyissä käytettäviä operaattoreita. #Sum-operaattori laskee hakuavainten painojen keskiarvon. #N-operaattori on läheisyysoperaattori, joka määrittelee, että hakuavainten on sijaittava kyselyyn täsmäävässä dokumentissa läheisyysoperaattorin määrittelemässä järjestyksessä siten, ettei vierekkäisten hakuavainten välissä ole n:ää useampaa sanaa. (Callan ym. 1992, 6–7.) #Uwn-operaattori on myös läheisyysoperaattori, mutta sen toimintaperiaate on hieman erilainen. Se nimittäin sallii hakuavainten sijaita löydettävässä dokumentissa missä tahansa järjestyksestä, kunhan vierekkäisten hakuavainten välissä ei ole n:ää useampaa kirjainta. #Wsum-operaattori määrittelee jotkut hakuavaimet tärkeämmiksi kuin toiset painottamalla vain joitakin avaimia. #Syn- eli synonyymioperaattori pitää hakuavaimia toistensa synonyymeina ja käsittelee kaikkia hakuavaimia samanarvoisina. (Rajashekar & Croft 1995, 275.) Tutkielmassa käytettävään Inqueryyn on lisätty Tampereen yliopiston informaatiotutkimuksen laitoksella yhdyssanaoperaattori #0, joka määrittelee, että hakuavainten tulee sijaita samassa sanassa annetussa järjestyksessä, mutta hakuavaimen kanssa täsmäävän yhdyssanan alussa, välissä tai lopussa voi olla muitakin yhdysosia.

6.2 Tutkimuksen testikokoelmat

Testikokoelmat jäljittelevät operationaalisia tiedonhakuympäristöjä ja mahdollistavat erilaisten tiedonhakustrategioiden ja -menetelmien tutkimisen laboratorio-oloissa (Voorhees 2003, 3). Testikokoelma koostuu joukosta hakuaiheita, suuresta määrästä dokumentteja eli dokumenttikokoelmasta ja saantikannasta, jossa määritellään, kullekin hakuaiheelle kaikki sen kannalta relevantit dokumentit. Käytännössä tietokannat ovat usein niin laajoja, että tyydytään jonkinlaiseen otokseen relevanteista dokumenteista. Relevanssiarviot ovat se tekijä, joka tekee dokumenttikokoelmasta ja hakuaiheista testikokoelman (Voorhees 2003, 4). Hakuaiheet (topic) ovat kirjalliseen muotoon puettuja tiedontarvekuvauksia, joissa on määritelty, mitä aihetta käsittelevää tietoa halutaan löytää ja mitä löydettävältä dokumentilta edellytetään, jotta se katsotaan relevantiksi.

Tässä opinnäytetyössä käytetään kahta testikokoelmaa, joista toinen sisältää suomen- ja toinen englanninkielisiä dokumentteja. Suomenkielinen TUTK-testikokoelma sisältää kokotekstimuotoisia sanomalehtiartikkeleita, jotka on julkaistu kolmessa suomalaisessa sanomalehdessä (Aamulehti, Keski-

suomalainen ja Kauppalehti) vuosina 1988–1992. Tietokanta sisältää lähestulkoon 54000 artikkelia. Keskimäärin artikkelit ovat melko lyhyitä, sillä ne ovat keskimäärin 202 sanaa pitkiä. (Sormunen 2000, 59.) Hakuaiheita on testikokoelmaa varten olemassa 35 kappaletta, joista tässä tutkimuksessa käytetään 30:a. TUTKissa relevanssiarviot on tehty dokumenteille neliportaista asteikkoa (asteikko 0-3) käyttäen. Arvioinnissa arvo 0 on annettu dokumenteille, jotka ovat täysin epärelevantteja, jotka eivät siis sisällä yhtään tietoa hakuaiheesta. Arvon 1 saaneet dokumentit ovat puolestaan marginaalisesti relevantteja, jolloin dokumentit ainoastaan viittaavat haettavaan aiheeseen, mutta eivät sisällä tietoa hakuaiheesta sen enempää kuin hakuaiheen kuvaus. Arvioinnissa arvo 2 on annettu dokumenteille, jotka katsotaan relevanteiksi ja jotka sisältävät hakuaiheesta muutamia uusia faktoja. Tämän arvon saanut dokumentti käsittelee yleensä vain joitakin aiheen teemoista. Arvon 3 saaneet dokumentit ovat erittäin relevantteja, koska ne sisältävät arvokasta tietoa aiheesta ja käsittelevät pääasiallisesti vain haettua aihetta. Tällaisessa dokumentissa käsitellään kaikkia tai lähes kaikkia aiheen alateemoja. (Sormunen 2000, 63; 2002, 325.)

Englanninkielisenä aineistona on TRECin (Text Retrieval Conference) kahden peräkkäisen, seitsemännen ja kahdeksannen, testikierroksen hakuaiheista, dokumenteista ja niihin liittyvistä relevanssiarvioista koostuva testikokoelma. Se valittiin käyttöön, koska siihen on jälkikäteen tehty Tampereella neliportaiset relevanssiarviot, joilla on korvattu alkuperäinen TRECin binäärinen relevanssiasteikko (Sormunen 2002, 324). Näin TRECin ja TUTKin relevanssiarvioista on saatu yhteensopivat. TREC on taho, joka tukee ja edistää tekstitiedonhakua tarjoamalla testikokoelmansa tiedonhakua tutkivan yhteisön käyttöön. TREC-testikokoelma on syntynyt testikierrosten tuloksena. Vuosittain järjestetään TREC-testikierros, jota varten testikokoelmaan laaditaan uusia hakuaiheita. Tiedonhakuyhteisö ottaa osaa testikierrokseen etsimällä hakuaiheisiin vastaavia dokumentteja omia tiedonhakujärjestelmiään käyttäen ja palauttaen saamansa tulokset TRECin arvioitavaksi. Testikierroksen päätteeksi järjestetään konferenssi (Text Retrieval Conference), jossa osanottajat jakavat kokemuksiaan. (TREC 2004.)

Pääasiallisesti TREC-testikokoelma koostuu sanomalehti- ja uutispalveluiden artikkeleista, mutta mukana on myös tietotekniikan alan tiivistelmiä sekä hallinnon toiminnan tuottamia aineistoja kuten patenttihakemuksia (Voorhees 2003, 3). Tässä käytetty testikokoelma sisältää noin 530 000 englanninkielistä artikkelia ja 41 hakuaihetta, joille on tehty neliportaiset relevanssiarviot. Arviointiasteikko on 0-3, jolloin asteikko alkaa arvosta 0 (=epärelevantti) ja päättyy arvoon 3 (=erittäin relevantti). Pe-

rusteet sille, minkä arvon dokumentti saa, ovat samat kuin TUTKissa. Arvioinnissa arvo 0 on annettu dokumenteille, jotka ovat täysin epärelevantteja. Arvon 1 saaneet dokumentit ovat puolestaan marginaalisesti relevantteja. Arvioinnissa arvo 2 on annettu dokumenteille, jotka katsotaan relevanteiksi. Arvon 3 saaneet dokumentit ovat erittäin relevantteja. (Sormunen 2000, 63; 2002, 325.)

Käytännössä tutkielman eräajoja tehtäessä relevanssiarviot jaettiin kummassakin testikokoelmassa kolmeen relevanssitasoon: liberaaliin, normaaliin ja tiukaan. Liberaalilla relevanssitasolla relevanteiksi katsottiin arvon 1, 2 tai 3 saaneet dokumentit. Normaalilla relevanssitasolla relevanteiksi katsottiin vain arvon 2 tai 3 saaneet dokumentit samalla kun arvon 0 tai 1 saaneita dokumentteja pidettiin epärelevantteina. Tiukalla relevanssitasolla relevantteina pidettiin vain arvon 3 saaneita dokumentteja ja muun arvon saaneita dokumentteja pidettiin epärelevantteina.

Englannin- ja suomenkielisten kyselyiden ja hakemistojen perusmuotoistamiseen käytettiin suomalaisen Lingsoft Oy:n ohjelmistoja. Englanninkielisten kyselyiden perusmuotoistamiseen käytettiin Engtwolia ja suomenkielisten kyselyiden perusmuotoistamiseen vastaavasti Fintwolia. Perusmuotoohjelmien toimintaperiaatteita on kuvattu tarkemmin luvussa 4.3. Englanninkielisten kyselyjen karsintaan käytettiin Porter-algoritmia ja suomenkieliset kyselyt karsittiin Snowball-ohjelmistolla. Kyseiset ohjelmistot on esitelty luvussa 4.2.3.

6.3 Tutkimuksen kyselysarjat ja eräajot

Tutkimuksen kyselyitä käsiteltiin kyselysarjoina. Yhden suomenkielisen kyselysarjan muodostivat TUTKin 30 kyselyä ja vastaavasti yhden englanninkielisen kyselysarjan TRECin 41 kyselyä. Silloin kyselyitä ei tarvinnut käsitellä yksi kerrallaan, vaan ne voitiin syöttää kyselysarjana esimerkiksi tiedonhakujärjestelmälle. Myös kaikki kyselyille tehtävät käsittelyt voitiin tehdä kerralla koko kyselysarjalle. Manuaalisessa käsittelyssä toimenpiteet täytyi tehdä yksitellen kullekin kyselylle.

Suomenkielisiä kyselysarjoja tehtiin kolme kappaletta: perusmuotoinen, karsittu sekä ositettu perusmuotoinen kyselysarja, joilla haettiin vastaavista tietokannan hakemistoista: perusmuotoisesta, karsitusta ja ositetusta perusmuotohakemistosta. Tarkoituksena oli verrata TUTKin perusmuotoisten kyselyiden tulosjoukkoja TUTKin karsittujen kyselyiden tulosjoukkojen kanssa. Lisäksi TUTKin ositettujen perusmuotoisten kyselyiden tulosjoukkoja verrattiin ensinnäkin TUTKin perusmuotoisten kyse-

lyiden ja toisekseen TUTKin karsittujen kyselyiden tulosjoukkojen kanssa. Ositetulle perusmuotoiselle kyselysarjalle ei ollut mahdollista laatia ositettua karsittua vastinetta, koska käytettävissä ei ollut keinoja ositetun karsitun hakemiston luomiseen. Ositettu perusmuotoinen kyselysarja tehtiin ikään kuin ylimääräisenä, jotta nähtäisiin osituksen vaikutus tuloksiin ja saataisiin tietää, onko sillä päällekkäisyyden määrään jokin lisävaikutus. Englanninkielisiä kyselysarjoja laadittiin vain kaksi: perusmuotoinen ja karsittu. Tarkoituksena oli verrata TRECin perusmuotoisten kyselyiden tulosjoukkoja TRECin karsittujen kyselyiden tulosjoukkojen kanssa. Ositettua kyselysarjaa ei laadittu, koska englannin kielen vähäisestä yhdyssanojen määrästä johtuen tutkimusaineistossa oli hyvin vähän sanoja, jotka olisi ylipäätään voitu osittaa. Tavallaan englanninkieliset kyselysarjat olivat jo luonnostaan ositettuja. Yhteensä kyselysarjoja laadittiin siis viisi kappaletta (ks. Taulukko 4). Taulukossa 5 havainnollistetaan myös kyselysarjojen välisiä pareittaisia vertailuja, joita tehtiin yhteensä neljä.

TAULUKKO 4. Tutkimuksen kyselysarjat.

Suomenkielinen testikokoelma TUTK:	**Englanninkielinen testikokoelma TREC:**
Perusmuotoinen kyselysarja	Perusmuotoinen kyselysarja
Ositettu perusmuotoinen kyselysarja	-
Karsittu kyselysarja	Karsittu kyselysarja

TAULUKKO 5. Kyselysarjojen keskinäiset vertailut.

TUTK:	Perusmuotoinen kyselysarja	vs.	Karsittu kyselysarja
	Ositettu perusmuotoinen kyselysarja	vs.	Karsittu kyselysarja
	Ositettu perusmuotoinen kyselysarja	vs.	Perusmuotoinen kyselysarja
TREC:	Perusmuotoinen kyselysarja	vs.	Karsittu kyselysarja

Seuraavassa esitellään yksityiskohtia kyselysarjojen sisältämien kyselyiden laadinnasta. Kyselyt oli tarpeen käsitellä hakemistojen käsittelyjä vastaaviksi. Kyselyt muodostettiin hakuaiheista mahdollisimman paljon tietotekniikan suomia automaattisia keinoja käyttämällä. Manuaalista käsittelyä tehtiin vain, jos sopivaa ohjelmaa ei ollut tarjolla. Tämä jäljittelee tiedonhakutapaa, jossa käyttäjän tehtäväksi jää vain hakukäsitteiden syöttäminen tiedonhakujärjestelmän hoitaessa hakemiston ja kyselyn hakuavainten käsittelyn automaattisesti käyttäjän puolesta ja hänen huomaamattaan. Testikokoelmien hakuaiheista tuli siis laatia varsinaiset tiedonhakujärjestelmälle syötettävät kyselyt. TUTK-

testikokoelmassa hakuaihekuvaus on yksiosainen ja se otettiin sellaisenaan kyselyiden laatimisen pohjaksi. Sen sijaan TRECin hakuaihekuvaukset koostuvat neljästä eri osiosta: identifioijasta (identifier), otsikosta (title), kuvauksesta (description) ja narratiivista (narrative) (Voorhees 2003, 3). Kuvio 2 sisältää esimerkin TRECin hakuaihekuvauksesta. Niinpä oli tarpeen ensin päättää, mistä TRECin hakuaihekuvauksen osiosta englanninkielisiä kyselyitä lähdettäisiin laatimaan.

```
<top>

<num> Number: 353
<title> Antarctica exploration

<desc> Description:
Identify systematic explorations and scientific investigations
of Antarctica, current or planned.

<narr> Narrative:
Documents discussing the following issues are relevant:

 - systematic explorations and scientific investigations of Antarctica
   (e.g., seismology, ionospheric physics, possible economic development)
 - other research currently conducted or planned for the future
 - banning of mineral mining

Documents discussing tourism are non-relevant.  Documents discussing
"disrupting scientific experiments" are non-relevant unless a specific
experiment is identified.

</top>
```

KUVIO 2. Esimerkki TRECin hakuaihekuvauksesta.

Identifioijalla tarkoitetaan kyselyn identifioivaa tunnistenumeroa. Otsikko osio sisältää enintään kolme aihetta keskeisimmin kuvaavaa sanaa. Kuvaus osio on yhden lauseen pituinen kuvaus etsittävästä aihealueesta. Narratiivissa kuvataan tiiviisti, mikä tekee löydetystä dokumentista kunkin aiheen kannalta relevantin. (Voorhees 2003, 3.) Käsittelyyn valittiin TRECin kuvaus (description) osio, sillä se oli lähimpänä TUTKin hakuaiheen laajuutta.

Kyselyiden laadinnassa oli useita eri vaiheita ja laadinta toteutettiin erilaisia ohjelmia käyttämällä. Ensimmäinen askel kyselyiden työstämisessä oli välimerkkien ja isojen alkukirjainten poistaminen. Se oli tarpeen, jotta muut kyselyjen työstämisessä käytettävät ohjelmat kykenisivät toimimaan virheettömästi. Toinen vaihe oli sulkusanojen poisto. Siinä ohjelma kävi läpi kyselyitä ja poisti niistä sanat, jotka löytyivät sulkusanalistalta. Sulkusanojen automaattisen poiston lisäksi kyselyistä poistet-

tiin manuaalisesti erilaisia lyhenteitä (ym, jne) sekä sanasta erilleen joutuneita ja sulkusanalistalle tuntemattomia sijapäätteitä kuten hakuavaimesta USA:ssa irralleen joutunut sijapääte ssa, sillä ne olivat rinnastettavissa sulkusanoihin.

Välimerkkien, isojen alkukirjainten sekä sulkusanojen poisto tehtiin kaikille kyselysarjoille. Sen jälkeen käsittely eriytyi kullekin kyselysarjalle omanlaiseksi. Kun kaksi yhteistä vaihetta oli tehty, voitiin perusmuotoisen kyselysarjan laatimisessa edetä kyselyiden perusmuoto-ohjelman avulla tapahtuvaan perusmuotoistamiseen. Perusmuotoisten kyselyiden laadinnassa sanat, mukaan lukien yhdyssanat, palautettiin perusmuotoihinsa. Osittamattomia perusmuotoisia kyselyitä laadittaessa yhdyssanoja ei pääsääntöisesti käsitelty sen enempää. Poikkeuksen tähän tekivät yhdysviivalliset yhdyssanat kuten ey-hakemus, joiden osat oli tarpeen yhdistää toisiinsa yhdysviivan sijasta yhdyssanaoperaattorin #0 avulla. Yhdyssanaoperaattori määrittelee, että avainten tulee sijaita samassa sanassa annetussa järjestyksessä, mutta hakuavainten kanssa täsmäävän yhdyssanan alussa, välissä tai lopussa voi olla muitakin yhdysosia. Esimerkiksi hakuavaimen #0(voima laitos) kanssa täsmäytyvät sanat voimalaitos, turvevoimalaitos ja voimalaitospato, mutta eivät sanat koululaitos tai lihasvoima.

Lopuksi mainitaan vielä perusmuotoisten kyselyiden laadinnassa esiintyneitä ongelmia. Kuten jo luvussa 4.3 todettiin, eivät perusmuoto-ohjelmat kykene tunnistamaan kaikkia juoksevan tekstin sanoja. Aina kun perusmuoto-ohjelma törmäsi sanaan, jota se ei kyennyt tunnistamaan, se lisäsi sanan eteen @-merkin. Erityisiä ongelmia perusmuoto-ohjelmalle aiheuttivat erisnimet. Lisäksi perusmuotoistamisen yhteydessä muutamasta kyselystä hävisivät numerot, jotka kuitenkin palautettiin manuaalisesti takaisin kyselyihin, sillä kyseiset numerot olivat osa kyselyn sisältöä. Esimerkiksi viidettä ydinvoimalaa käsittelevässä kyselyssä numero 5 sanan ydinvoimala edessä oli tarkoituksenmukainen osa kyselyä.

Kahden ensimmäisen työvaiheen jälkeen ositetun perusmuotoisen kyselysarjan laatimisessa edettiin vaiheeseen, jossa perusmuoto-ohjelma teki perusmuotoistamisen ohella myös yhdyssanojen osituksen. Perusmuoto-ohjelma antoi ositusta suorittaessaan tulokseksi sekä ositetun yhdyssanan yhdysosat että samaisen yhdyssanan osittamattomana. Näin saadut yhdysosat nivottiin toisiinsa yhdyssanaoperaattorin avulla. Yhdyssanaoperaattorilla yhdistetyillä yhdysosilla kyetään löytämään samat ja jopa enemmän dokumentteja kuin osittamattomalla yhdyssanalla, joten itse yhdyssanan säilyttäminen kyselyssä olisi ollut täysin tarpeetonta toistoa. Niinpä ne voitiin poistaa kyselyistä manuaalisesti. Fint-

wol esitti yhdyssanojen osat osituksen jälkeen hieman takaperoisessa järjestyksessä, esimerkiksi ase, ydin, ydinase, jolloin yhdysosien järjestystä jouduttiin vaihtamaan keskenään manuaalisesti, jotta lopullisessa kyselyssä tuloksena olisi #0(ydin ase).

Kun isot kirjaimet, välimerkit ja sulkusanat oli poistettu, oli seuraava työvaihe karsitun kyselysarjan laatimisessa karsinta-algoritmilla karsiminen. Suomenkielisen aineiston käsittelemiseen käytettiin Snowball-ohjelmistoa ja englanninkielisen aineiston käsittelyyn Porter-algoritmia. Karsittujen kyselyiden yhdysviivallisten yhdyssanojen yhdistämiseen ei käytetty yhdyssanaoperaattoria vaan sanojen väliin jätettiin yhdysviiva. Ero perusmuotoisten ja karsittujen kyselyjen yhdysviivallisten sanojen käsittelyssä johtui kohdehakemistojen erilaisesta yhdysviivallisten sanojen käsittelystä hakemistojen laadintavaiheessa.

Lisäksi oli vielä tarpeen yhtenäistää toisiinsa verrattavia suomenkielisiä kyselyitä. Kuten tiedetään karsinta-algoritmi ja perusmuoto-ohjelma toimivat eri tavoin, sillä karsinta-algoritmi karsii vain annetun sanan, kun taas perusmuoto-ohjelma saattaa tuottaa samasta sanasta monta luentaa, jolloin karsitut kyselyversiot jäävät verrokkeinaan toimivia perusmuotoisia kyselyversioita suppeammiksi. Käytettyjen ohjelmien erilaisesta toimintatavasta johtuen TUTKin kyselysarjat eivät vastanneet toisiaan laajuudeltaan. Karsitusta ja perusmuotoisesta kyselysarjasta oli tarpeen saada mahdollisimman samankaltaiset, jotta ne ja niillä saatavat tulokset olisivat keskenään mahdollisimman vertailukelpoisia. Ylimääräiset luennat olisivat antaneet perusmuotoisille kyselyille enemmän painoarvoa, jolloin ne olisivat saattaneet menestyä hauissa paremmin. Kyselyitä piti siis yhdenmukaistaa poistamalla perusmuotoisista kyselyistä perusmuoto-ohjelman tuottamat ylimääräiset luennat. Kyselyissä säilyivät vain semanttisesti mielekkäät luennat. Saman ongelman välttämiseksi myös ositetusta perusmuotoisesta kyselysarjasta poistettiin nämä ylimääräiset luennat, jotta se olisi laajuudeltaan mahdollisimman samankaltainen osittamattoman perusmuotoisen kyselysarjan kanssa. Tämä ongelma ei koskettanut englanninkielisiä kyselysarjoja, sillä perusmuoto-ohjelma ei ollut lisännyt niihin ylimääräisiä luentoja.

Lopuksi kaikille kyselysarjoille yhteinen toimenpide oli kyselyiden varustaminen käytettävän tiedonhakujärjestelmän vaatimilla operaattoreilla ja muilla sen kyselyn syntaksilta vaatimilla merkeillä. Tarkoituksena oli laatia rakenteettomia kyselyitä, koska ei haluttu kyselyn rakenteen vaikuttavan saataviin tuloksiin. Niinpä kyselyissä käytettiin Inqueryn #sum-operaattoria, joka painottaa kaikkia ha-

kuavaimia yhtä paljon ja laskee avainten painoarvoista keskiarvon. Operaattoriksi valittiin #sum-operaattori, koska Sormusen, Kekäläisen, Koiviston ja Järvelinin (2001, 363) mukaan #sum-operaattori on paras rakenteettomien kyselyiden laatimisessa käytettävissä oleva operaattori.

```
#q4=#sum(#0(jyväs kylä) kaupunki #0(maalais kunta) #0(kunta liitos hanke) kartoittaa #0(liitos hanke)
kannattaja vastustaja #0(mieli pide) perustelu arvio liitos taloudellinen vaikutus #0(porkkana raha));
```

KUVIO 3. Esimerkki Inqueryn syntaksilla varustetusta ositetusta kyselystä.

Kun kyselyt oli saatu täysin valmiiksi, ne syötettiin tiedonhakujärjestelmälle kyselysarjoittain. Tätä voidaan kutsua myös eräajojen suorittamiseksi. Tuloksena tiedonhakujärjestelmä palautti kunkin kyselyn tulosjoukon. Sen perusteella saatiin tietää, montako relevanttia dokumenttia tulosjoukko sisältää. Lisäksi tiedonhakujärjestelmä ilmoitti jokaisen kyselyn tarkkuusluvut saantitasoittain sekä keskitarkkuudet. Sananmuotojen käsittelymenetelmien tuloksellisuuden tarkastelu perustui näihin tietoihin. Tiedonhakujärjestelmä kertoi myös, minkä numeroinen dokumentti löytyi mistäkin tulosjoukon positiosta. Viimeksi mainittuja tietoja käytettiin tulosjoukkojen päällekkäisyyksiä tarkasteltaessa.

6.4 Tulosjoukkojen päällekkäisyyden vertailun periaate

Eräajojen suorittamisen jälkeen tutkimuksessa voitiin edetä tarkastelemaan tulosjoukkojen päällekkäisyyttä. Ennen kuin kerrotaan varsinaisia tulosjoukkojen vertailun tuloksia, kuvataan niitä periaatteita, joita noudattaen tulosjoukkojen päällekkäisyyden tarkastelu tehtiin. Päällekkäisyyden tutkimiseen liittyviä periaatteellisia asioita ovat muun muassa se, miten päällekkäisyys ymmärretään toisaalta täystäsmäyttävässä ja toisaalta osittaistäsmäyttävässä järjestelmässä.

6.4.1 Päällekkäisyyden tutkimisen periaate täys- ja osittaistäsmäyttävässä tiedonhakujärjestelmässä

Tulosjoukkojen päällekkäisyyden vertailu ymmärretään eri tavalla riippuen siitä, onko käytössä täys- vai osittaistäsmäyttävä tiedonhakujärjestelmä. Niinpä luvuissa 5.2 ja 5.3 esitellyissä tutkimuksissa päällekkäisyyttä on tarkasteltu keskenään eri tavalla. Täystäsmäyttävää tiedonhakujärjestelmää käytettäessä voidaan toisiinsa vertailla kokonaisia tulosjoukkoja. Osittaistäsmäyttävässä järjestelmässä

tulosjoukko katkaistaan halutun kokoiseksi tulosjoukkojen vertailua varten. Sitä kohtaa, josta tuo katkaisu tehdään, kutsutaan katkaisupisteeksi. Niinpä osittaistäsmäyttävässä järjestelmässä tulosjoukon koko määräytyy kulloisenkin katkaisupisteen mukaan. Nimitys katkaisupiste on syytä vaihtaa siinä vaiheessa, kun kahta samasta katkaisupisteestä katkaistua tulosjoukkoa aletaan verrata toisiinsa. Silloin on parempi puhua katkaisupisteen sijaan vertailupisteestä, sillä tarkastelun näkökulma muuttuu. Enää ei olekaan kyse kohdasta, josta tulosjoukko katkaistaan vaan kohdasta, jossa vertailu tehdään. Koska tulosjoukon koko määräytyy eri tavalla osittais- ja täystäsmäyttävässä järjestelmässä, tarkastellaan päällekkäisyyttä osittaistäsmäyttävää järjestelmää käytettäessä yleensä useissa eri vertailupisteissä.

Päällekkäisyydellä tarkoitetaan osittaistäsmäyttävässä tiedonhaussa tulosjoukkojen yhteisten dokumenttien suhteellista määrää tietyssä vertailupisteessä, mikä tarkoittaa että tulosjoukkojen yhteisten dokumenttien määrä suhteutetaan kulloiseenkin vertailupisteeseen. Jotta osittaistäsmäyttävässä järjestelmässä tulosjoukkojen dokumenttien käytännössä katsotaan olevan päällekkäisiä, dokumenttien tulee sisältyä saman vertailupisteen rajaamaan dokumenttijoukkoon, mutta niiden ei tarvitse sijaita vertailtavissa tulosjoukoissa identtisissä positioissa. Jotta päinvastaisessa tilanteessa voitaisiin sanoa, etteivät vertailtavien tulosjoukkojen dokumentit ole päällekkäisiä, dokumentit eivät saa sisältyä saman vertailupisteen rajaamaan dokumenttijoukkoon. Toisin sanoen osittaistäsmäyttävässä järjestelmässä dokumenttien positiot ovat merkityksellisiä, kun tulosjoukkoja verrataan toisiinsa. Täystäsmäyttävä järjestelmä eroaa osittaistäsmäyttävästä järjestelmästä tässä suhteessa, sillä täystäsmäyttävässä järjestelmässä koko tulosjoukkoa pidetään relevanttina, jolloin verrattavien tulosjoukkojen dokumenttien positioilla ei ole väliä. Täys- ja osittaistäsmäyttävä järjestelmä eroaa toisistaan vielä siinäkin mielessä, että täystäsmäyttävällä järjestelmällä saatuja tulosjoukkoja toisiinsa vertailtaessa saattavat vertailtavat tulosjoukot olla keskenään hyvin erikokoisia, mikä saattaa vääristää päällekkäisyyden määrää. Osittaistäsmäyttävää järjestelmää käytettäessä vertailtavat tulosjoukon osat ovat sen sijaan keskenään samankokoisia.

Edellä siis todettiin, että täystäsmäyttävää tiedonhakujärjestelmää käytettäessä voidaan toisiinsa vertailla kokonaisia tulosjoukkoja. Samalla todettiin, että osittaistäsmäyttävässä järjestelmässä tulosjoukko katkaistaan halutun kokoiseksi tulosjoukkojen vertailua varten. Tästä huolimatta puhe kokonaisista tulosjoukoista ei rajoitu tässä työssä pelkästään täystäsmäyttävän tiedonhaun piiriin. Jatkossa tullaan puhumaan kokonaisista tulosjoukoista ja päällekkäisyydestä kokonaisissa tulosjoukoissa,

vaikka toimitaankin osittaistäsmäyttävän järjestelmän kontekstissa. Tällöin puhe kokonaisista tulosjoukoista tulee kuitenkin ymmärtää vastakohtana tilanteelle, jossa tarkastelu on rajoittunut pelkästään tulosjoukkojen relevantteihin osiin.

Kokonaisten tulosjoukkojen päällekkäisyyden tarkastelun periaate

Seuraavaksi esitetään periaate, jota noudattaen tässä tutkielmassa tarkasteltiin kokonaisten tulosjoukkojen päällekkäisyyttä. Eräajojen tuloksena saatujen tulosjoukkojen vertailussa toimittiin niin, että kahden verrattavan tulosjoukon yhteisten dokumenttien määrä suhteutettiin tulosjoukon kulloiseenkin vertailupisteeseen. Näitä tietoja käytettiin päällekkäisyyden tarkastelemiseen kokonaisissa tulosjoukoissa. Kun päällekkäisyyttä tarkastellaan kokonaisissa tulosjoukoissa, tarkastellaan kaikkia vertailtaviin tulosjoukkoihin päätyneitä dokumentteja, olivat ne sitten epärelevantteja tai relevantteja, jotta saataisiin selville, missä määrin ne osoittautuvat toisiinsa verrattaville tulosjoukoille yhteisiksi. Kokonaisten tulosjoukkojen päällekkäisyyttä tarkasteltiin yhteensä 8:ssa eri vertailupisteessä, joita olivat vertailupisteet: 5, 10, 20, 50, 100, 200, 500 ja 1000. Tällöin tulosjoukkojen yhteisten dokumenttien määrää tarkastellaan, kun on löydetty 5 dokumenttia, 10 dokumenttia ja niin edelleen. Alkupään vertailupisteet vertautuvat tilanteeseen, jossa käyttäjä jaksaa käydä läpi vain muutaman tai muutaman kymmentä dokumenttia molempien tulosjoukkojen alusta. Vertailupisteiden lisäksi päällekkäisyyttä tarkasteltiin eri relevanssitasoilla, joista on kerrottu tarkemmin luvussa 6.2.

Seuraavaksi esitetään kaava, jota tässä tutkielmassa on käytetty laskettaessa päällekkäisyys kokonaisissa tulosjoukoissa:

$$\text{Päällekkäisyys}_T = \frac{|T_1 \cap T_2|}{v}$$

T_i = tulosjoukko
v = vertailupiste

Kun päällekkäisyyttä tarkastellaan kokonaisissa tulosjoukoissa, tehdään tarkastelu suhteessa vertailupisteeseen, mikä tarkoittaa, että toisiinsa verrattavien tulosjoukkojen yhteisten dokumenttien määrä jaetaan kulloisellakin vertailupisteellä.

Myös sitä tarkasteltiin, miten paljon kokonaisille tulosjoukoille yhteisiksi osoittautuneista dokumenteista oli relevantteja ja miten paljon epärelevantteja dokumentteja. Epärelevanttien määrä saatiin selville vähentämällä kokonaisille tulosjoukoille yhteisten dokumenttien määrästä relevanttien dokumenttien määrä.

Tulosjoukkojen relevanttien osien päällekkäisyyden tarkastelun periaate

Kun tarkastellaan vain tulosjoukkojen relevantteja osia, verrataan kahden tulosjoukon relevantteja dokumentteja toisiinsa sen selvittämiseksi, miten paljon tulosjoukkojen välillä esiintyy yhteisiä relevantteja dokumentteja. Myös tässä tarkastelussa käytettiin 8:aa eri vertailupistettä, jotka mainittiin kertaalleen jo edellisessä alaluvussa.

Kaava, jolla tulosjoukkojen relevanttien osien päällekkäisyys on laskettu tässä tutkielmassa, on seuraavanlainen:

$$\text{Päällekkäisyys}_{R} = \frac{|R_1 \cap R_2|}{r}$$

R_i = Tulosjoukon T_i relevanttien dokumenttien joukko

r= min ($|R_1 \cup R_2|$, v)

Kun tarkastellaan tulosjoukkojen relevanttien osien päällekkäisyyttä kussakin vertailupisteessä, jaetaan verrattavien tulosjoukkojen yhteisten relevanttien dokumenttien määrä verrattavien tulosjoukkojen kaikkien relevanttien dokumenttien määrällä tai kyseisellä vertailupisteellä, mikäli kaikkien relevanttien määrä on suurempi kuin kyseessä oleva vertailupiste.

7 Tulokset

7.1 Eri kyselyversioiden tarkkuuden interpoloidut keskiarvot

Ennen kuin tarkastellaan tulosjoukkojen päällekkäisyyksiä, katsotaan millainen on karsinnan, perusmuotoistamisen ja osittamisen tuloksellisuus tässä tutkimuksessa. Tuloksellisuuden tarkasteleminen on paikallaan, sillä se luo pohjan päällekkäisyyksien tarkastelulle. Jo aiemmissa päällekkäisyyttä tarkastelleissa tutkimuksissa on todettu, että menetelmät, joilla saadut tulosjoukot ovat olleet vain vähäisessä määrin päällekkäisiä, ovat olleet tuloksellisuudeltaan samankaltaisia (mm. Katzer ym. 1982, 272; McGill ym. 1979, 75–76). Niinpä tässäkin tutkimuksessa on tarpeen katsoa, ovatko menetelmät tuloksellisuudeltaan melko samanlaisia. Tuloksellisuutta päästään tarkastelemaan Inqueryn antamien saanti-tarkkuusarvojen avulla. Ensin katsotaan tarkkuuksia englanninkielisessä aineistossa.

7.1.1 Tarkkuudet englanninkielisessä aineistossa

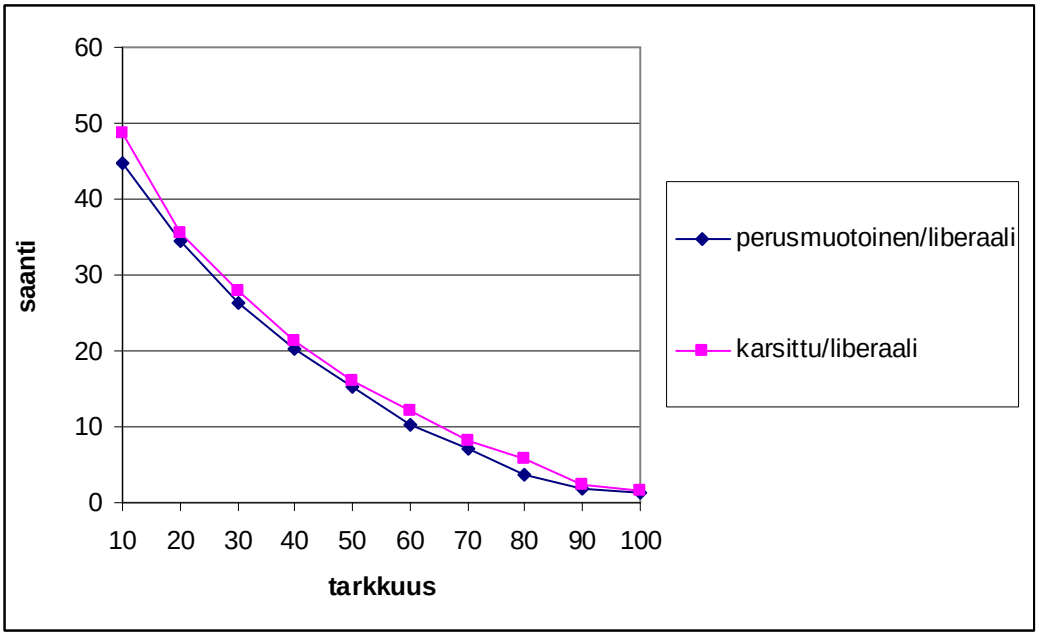
Kun karsittujen kyselyiden tarkkuuskeskiarvoja verrataan perusmuotoisilla kyselyillä saatuihin tarkkuuskeskiarvoihin toisiaan vastaavilla relevanssitasoilla, havaitaan, että englanninkielisellä aineistolla parhaat tarkkuuskeskiarvot saadaan karsituilla kyselyillä (ks. Taulukko 6). Karsitun kyselysarjan tarkkuuskeskiarvot ovat 1-1,5 prosenttiyksikön verran parempia kuin perusmuotoisen kyselysarjan. Kuvioissa 4, 5 ja 6 esitetään perusmuotoisten ja karsittujen kyselyjen saanti-tarkkuuskäyrät liberaalilla, normaalilla ja tiukalla relevanssitasolla. Kuvioista nähdään, että karsintaa ja perusmuotoistamista kuvaavat käyrät menevät hyvin läheltä toisiaan niin että karsintaa kuvaavat käyrät ovat vain hieman perusmuotoistamista kuvaavia käyriä ylempänä. Tässä esitettyjen tarkkuuskeskiarvojen ja saanti-tarkkuuskäyrien perusteella voidaan todeta, ettei karsinnan ja perusmuotoistamisen tuloksellisuudessa ole juurikaan eroja englanninkielistä aineistoa käytettäessä. Tältä osin tämän tutkimuksen tulokset ovat linjassa aiempien tutkimusten tulosten kanssa.

Englanninkielisen aineiston tarkkuuskeskiarvot ovat jossain määrin yllättäviä. Tavallisesti tuloksellisuus laskee johdonmukaisesti relevanssitasoittain, koska tietokannan saantikanta pienenee relevanssitasoittain siten että liberaalilla relevanssitasolla saantikanta sisältää eniten relevantteja dokumentteja, normaalilla relevanssitasolla toiseksi eniten ja tiukalla relevanssitasolla vähiten relevantteja dokumentteja. Kun saantikannan koko pienentyy jokaisella relevanssitasolla, muuttuu kyselyjen täsmäy-

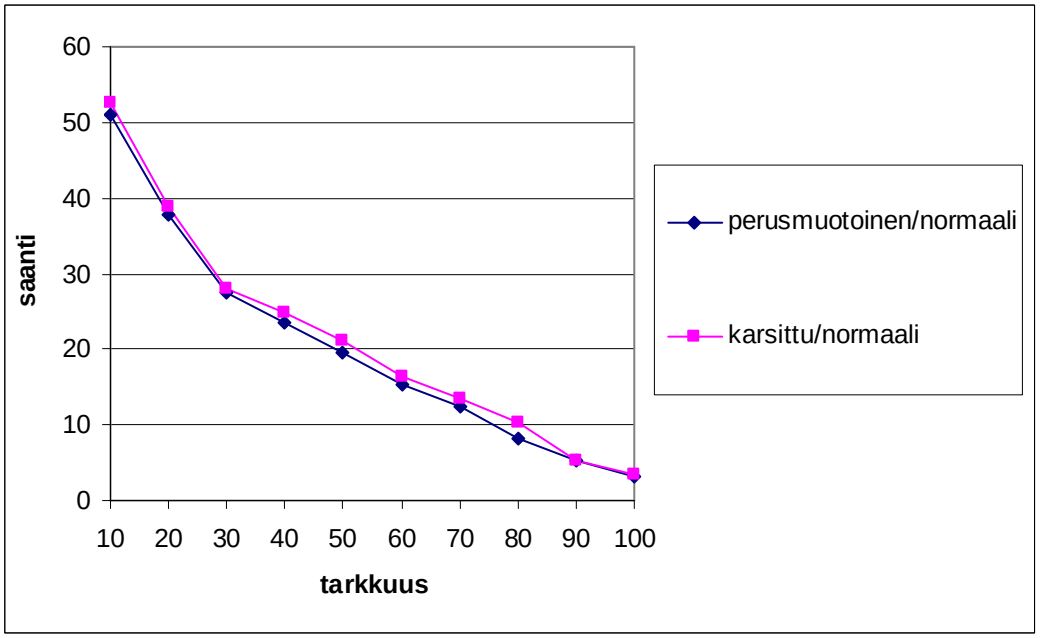
tyminen relevantteihin dokumentteihin vaikeammaksi, koska relevantteja dokumentteja, joihin kysely voisi täsmäytyä, on olemassa aina vain vähemmän. Tällöin seurauksena on, että tulosjoukkojen relevanttien dokumenttien määrä ja sitä kautta kyselyjen tarkkuudet pienenevät relevanssitasoittain. Englanninkielisen aineiston tarkkuuskeskiarvot ovat yllättäviä siksi, että keskitarkkuudet käyttäytyvät vastoin äsken esitettyä kuvausta, sillä keskitarkkuus on korkein normaalilla relevanssitasolla, toiseksi korkein tiukalla relevanssitasolla ja matalin liberaalilla relevanssitasolla (ks. Taulukko 6). Tämän yllättävän havainnon selittämiseksi englanninkielisten perusmuotoisten ja karsittujen kyselysarjojen saanti-tarkkuusarvoja tarkastellaan vielä lisää.

TAULUKKO 6. (TREC) Perusmuotoisen ja karsitun kyselysarjan tarkkuuden interpoloidut keskiarvot eri relevanssitasoilla englanninkielisessä aineistossa.

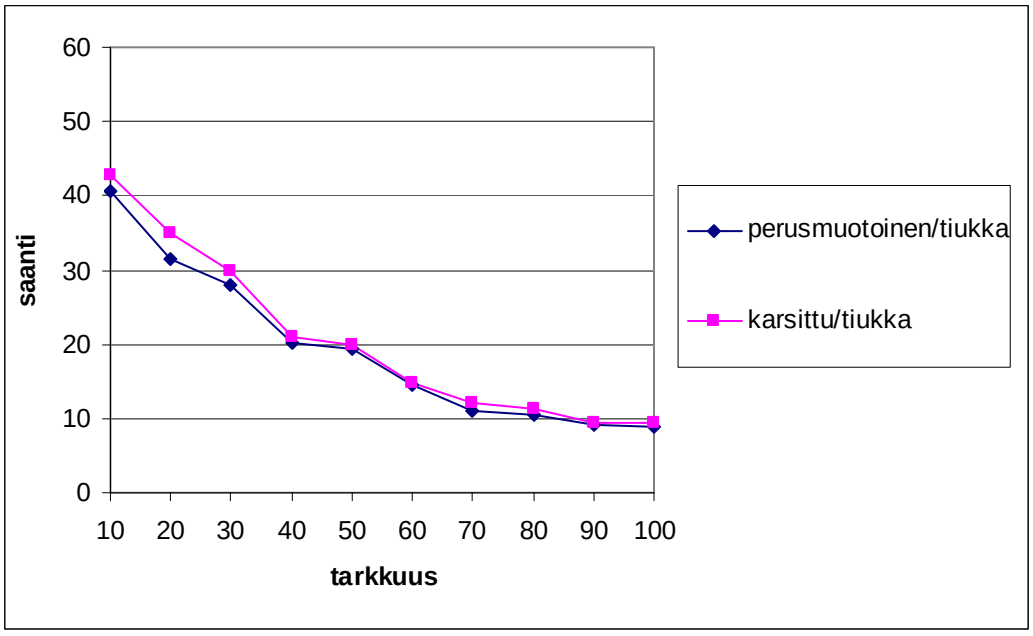
	Perusmuotoinen kyselysarja	Karsittu kyselysarja	Ero prosenttiyksiköissä
Tarkkuuskeskiarvo:			
Liberaali relevanssitaso	16,5	18	1,5
Normaali relevanssitaso	20,4	21,4	1,0
Tiukka relevanssitaso	19,3	20,6	1,3



KUVIO 4. (TREC) Perusmuotoisten ja karsittujen kyselyjen saanti-tarkkuuskäyrät liberaalilla relevanssitasolla englanninkielisessä aineistossa.



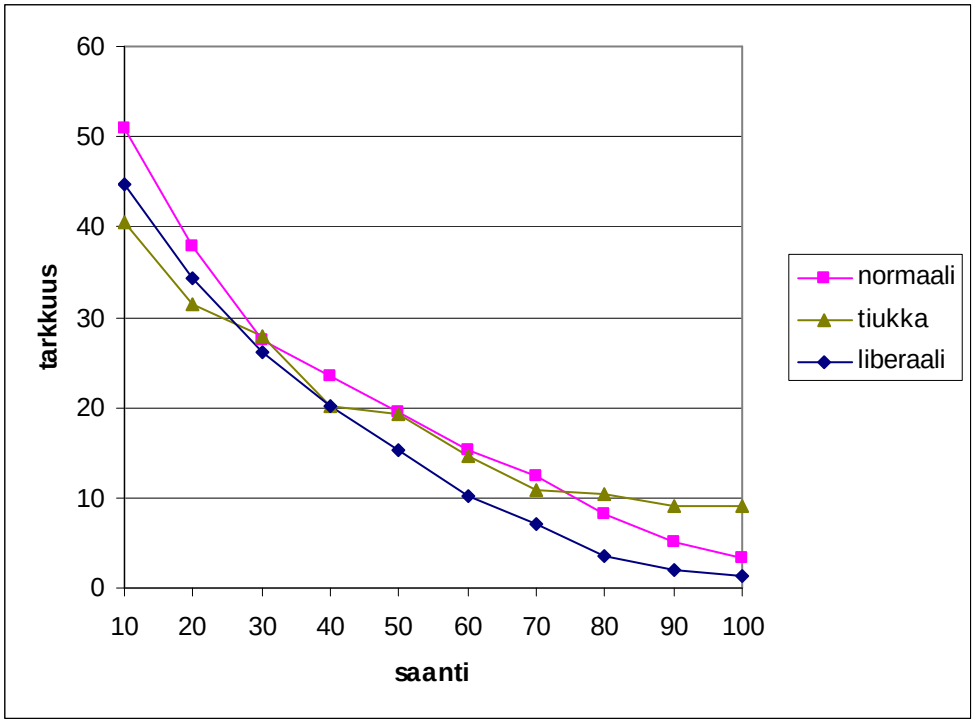
KUVIO 5. (TREC) Perusmuotoisten ja karsittujen kyselyjen saanti-tarkkuuskäyrät normaalilla relevanssitasolla englanninkielisessä aineistossa.



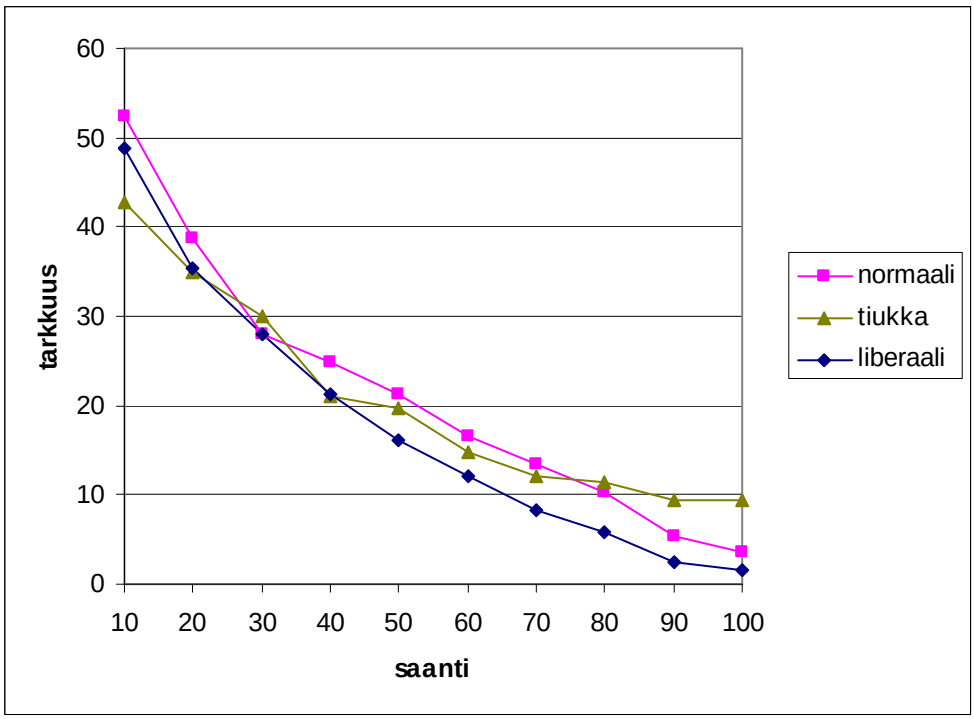
KUVIO 6. (TREC) Perusmuotoisten ja karsittujen kyselyjen saanti-tarkkuuskäyrät tiukalla relevanssitasolla englanninkielisessä aineistossa.

Kuviot 7 ja 8 havainnollistavat perusmuotoisilla ja karsituilla kyselyillä saatujen saanti-tarkkuusarvojen yllättävää käyttäytymistä. Saanti-tarkkuuskäyristä nähdään, että tarkkuus on suurinta vuorotellen joko normaalilla tai tiukalla relevanssitasolla. Käytetäänpä sitten karsittuja tai perusmuotoisia kyselyitä on tarkkuus suurinta normaalilla relevanssitasolla saantitasolle 20 asti eli aina siihen asti että 20 prosenttia tietokannan relevanteista dokumenteista on löydetty. Saantitasolla 30 tarkkuus on sen sijaan suurinta tiukalla relevanssitasolla. Tarkat TRECin saanti-tarkkuusarvot löytyvät tutkielman liitteistä (ks. Liite 1). Saantitason 30 jälkeen normaalin relevanssitason tarkkuudet nousevat jälleen suurimmiksi, sillä tarkkuus on suurinta normaalilla relevanssitasolla saantitasoilla 40, 50, 60 ja 70. Saantitasoilla 80, 90 ja 100 tarkkuus on jälleen suurinta tiukalla relevanssitasolla. Tarkkuus liberaalilla relevanssitasolla vaihtelee eri tavoin riippuen siitä, tarkastellaanko perusmuotoisia vai karsittuja kyselyitä. Kuviosta 7 nähdään tarkkuuden olevan perusmuotoisia kyselyitä käytettäessä matalin liberaalilla relevanssitasolla saantitasolla 30 ja siitä eteenpäin aina saantitasolle 100 asti. Ennen saantitasoa 30 liberaalin relevanssitason tarkkuudet ovat toiseksi parhaat. Kuviosta 8 nähdään, että karsittuja kyselyitä käytettäessä tarkkuus on alhaisin liberaalilla relevanssitasolla saantitason 40 jälkeen, jollaisena se pysyy aina saantitasolle 100 asti. Karsittuja kyselyitä käytettäessä liberaalin relevanssitason tarkkuus menee saantitasoilla 30 ja 40 tasan joko normaalin tai tiukan relevanssitason tarkkuuksien kanssa. Ennen saantitasoa 30 liberaalin relevanssitason tarkkuudet ovat toiseksi parhaat. Kuvioiden 7 ja 8 perusteella voidaan sanoa, että poikkeuksellinen tarkkuuskeskiarvojen käyttäytymi-

nen selittyy sillä, että relevantit ja erittäin relevantit löytyvät tulosjoukosta keskimäärin ennen marginaalisesti relevantteja, jotka sijoittuvat tulosjoukossa pääosin niiden jälkeen.



KUVIO 7. (TREC) Perusmuotoisilla kyselyillä saadut tarkkuudet eri relevanssitasoilla.



KUVIO 8. (TREC) Karsituilla kyselyillä saadut tarkkuudet eri relevanssitasoilla.

7.1.2 Tarkkuudet suomenkielisessä aineistossa

Suomenkielisessä aineistossa parhaat tarkkuuskeskiarvot saadaan ositetuilla perusmuotoisilla kyselyillä, toiseksi parhaat perusmuotoisilla kyselyillä ja heikoimmat karsituilla kyselyillä (ks. Taulukko 7). Kyselysarjojen keskitarkkuus laskee relevanssitasoittain siten, että keskitarkkuus on korkein liberaalilla ja matalin tiukalla relevanssitasolla. Tässä tutkimuksessa osittamisen, perusmuotoistamisen ja karsinnan paremmuusjärjestys on sama kuin Kettusella ym. (2005). Tämä tutkimus eroaa Kettusen ym. (2005) tutkimuksesta sen suhteen, että tässä ei käytetty lainkaan kyselyitä, joiden hakuavaimet olivat vartalo-ohjelmien tuottamia vartaloita. Mutta menetelmät, joita kummassakin tutkimuksessa käytettiin, menivät tuloksellisuuden perusteella tarkasteltaessa samaan paremmuusjärjestykseen. Taulukosta 7 nähdään, että ositetun perusmuotoisen ja perusmuotoisen kyselysarjan tarkkuuskeskiarvojen väliset erot ovat hyvin pienet, sillä ositetun perusmuotoisen kyselysarjan tarkkuuskeskiarvot ovat vain 0-0,8 prosenttiyksikköä paremmat kuin perusmuotoisen kyselysarjan tarkkuuskeskiarvot. Myös kuvioista 9, 10 ja 11 nähdään, että osittamisen ja osittamatta jättämisen käyrät menevät lähestulkoon päällekkäin. Ositetun perusmuotoisen kyselysarjan käyrä on hieman perusmuotoisen kyselysarjan käyrää ylempänä ainoastaan parin ensimmäisen saantitason kohdalla ja silloinkin vain normaalilla ja tiukalla

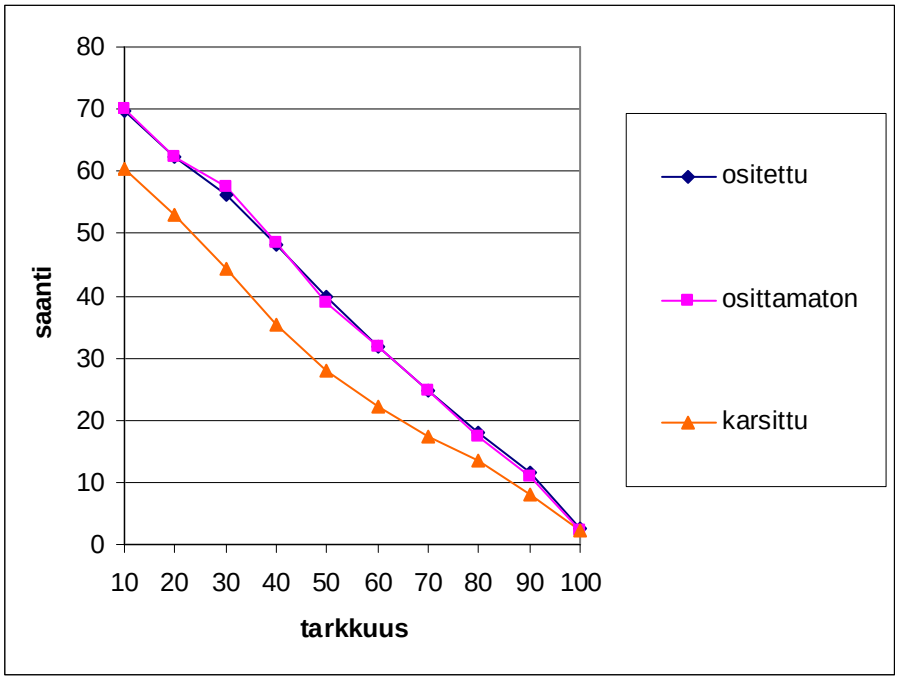
relevanssitasolla. Niinpä tämän tutkimuksen tulosten valossa ositetun perusmuotoisen ja perusmuotoisen kyselysarjan tuloksellisuudessa ei ole juuri lainkaan eroa toisiinsa nähden. Nämä havainnot ovat linjassa Kuntun (2003, 72) tutkimuksen kanssa siltä osin, että myös siinä osittaminen osoittautui paremmaksi avainten käsittelyyn käytettäväksi menetelmäksi kuin ilman ositusta tapahtuva perusmuotoistaminen. Erona on se, että Kuntun (2003, 51–58) tutkimuksessa näiden kahden menetelmän välinen ero oli suurempi, sillä siinä osittamisen ja osittamatta jättämisen tarkkuuskeskiarvot poikkesivat toisistaan relevanssitasosta riippuen 1,2–3 prosenttiyksikköä ositetun perusmuotoisen kyselysarjan eduksi.

TAULUKKO 7. (TUTK) Ositetun perusmuotoisen, perusmuotoisen ja karsitun kyselysarjan tarkkuuden interpoloidut keskiarvot eri relevanssitasoilla suomenkielisessä aineistossa.

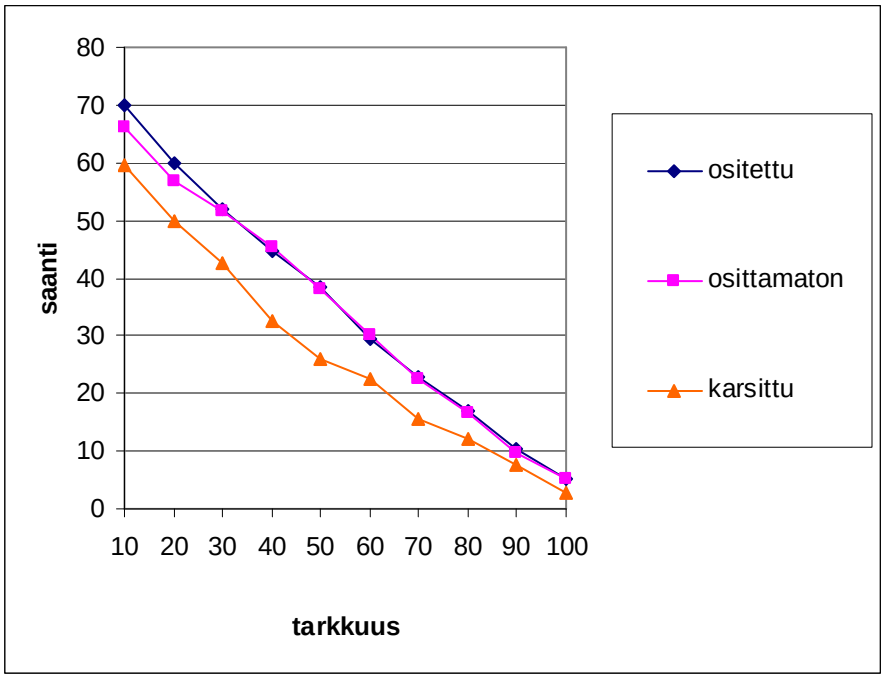
	Ositettu perusmuotoinen kyselysarja	Osittamaton perusmuotoinen kyselysarja	Ositetun ja osittamattoman ero prosenttiyksiköissä	Karsittu kyselysarja	Ositetun ja karsitun ero prosenttiyksiköissä
Tarkkuuskeskiarvo:					
Liberaali relevanssitaso	36,5	36,5	0	28,5	8
Normaali relevanssitaso	35	34,2	0,8	27,2	7,8
Tiukka relevanssitaso	24,5	24,3	0,2	20,4	4,1

TUTKissa perusmuotoisen ja karsitun kyselysarjan väliltä löytyvät erot ovat suurehkot, verrattiinpa karsitun kyselysarjan kanssa sitten ositettua perusmuotoista tai osittamatonta perusmuotoista kyselysarjaa. Ositetun perusmuotoisen ja karsitun kyselysarjan tarkkuuskeskiarvot poikkeavat toisistaan 4,1–8 prosenttiyksikön verran (ks. Taulukko 7) ja osittamattoman perusmuotoisen ja karsitun kyselysarjan tarkkuuskeskiarvot 3,9–8 prosenttiyksikön verran. Tilan niukkuuden vuoksi Taulukossa 7 esitetään vain ositetun perusmuotoisen ja karsitun kyselysarjan väliset erot. Tätä voidaan perustella sillä, että ositetun perusmuotoisen ja karsitun kyselysarjan välisen eron tiedetään osituksen ansiosta olevan suurempi kuin osittamattoman perusmuotoisen ja karsitun kyselysarjan välinen ero. Näin ollen eroa

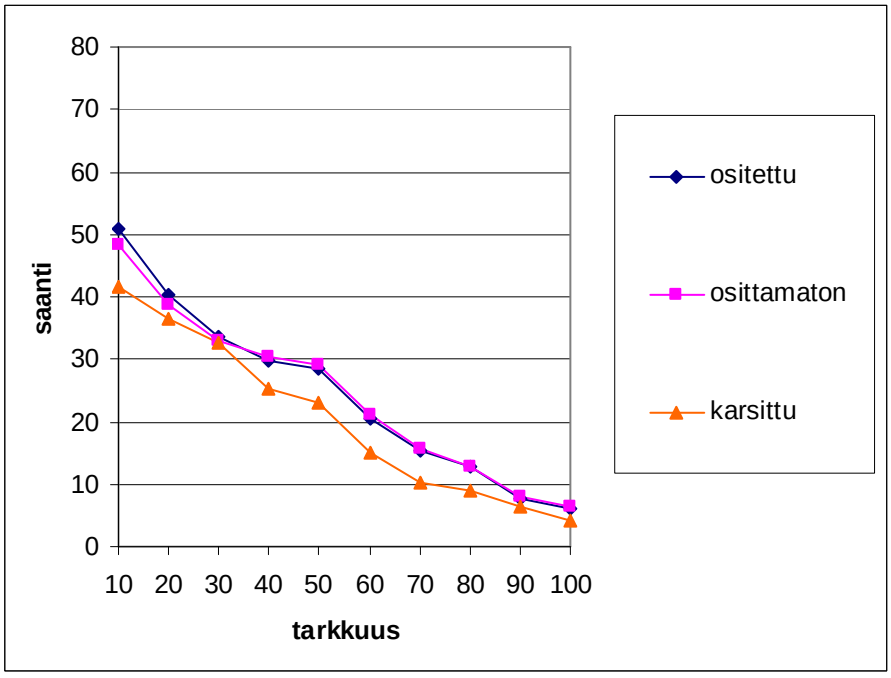
osittamattoman perusmuotoisen ja karsitun kyselysarjan välillä ei esitetä, koska se ei kertoisi enää mitään oleellisesti uutta tietoa. Taulukossa 7 esitettävien tietojen pohjalta voidaan kuitenkin laskea myös puuttuvan sarakkeen tiedot. Sparck Jonesin (1974, 397) peukalosäännön mukaan viiden – kymmenen prosenttiyksikön ero on jo käytännössä havaittava. Niinpä erot ositetun perusmuotoisen ja karsitun kyselysarjan välillä, sekä osittamattoman perusmuotoisen ja karsitun kyselysarjan välillä, ovat käytännössä havaittavia liberaalilla ja normaalilla relevanssitasolla. Myös kuvioista 9, 10 ja 11 nähdään, että karsitun kyselysarjan saanti-tarkkuuskäyrät ovat selvästi alempana kuin ositettujen perusmuotoisten ja osittamattomien perusmuotoisten kyselysarjojen käyrät. Karsitun kyselysarjan käyrät ovat huomattavasti paljon alempana kaikilla relevanssitasoilla.



KUVIO 9. (TUTK) Ositettujen perusmuotoisten, osittamattomien perusmuotoisten ja karsittujen kyselyjen saanti-tarkkuuskäyrät liberaalilla relevanssitasolla suomenkielisessä aineistossa.



KUVIO 10. (TUTK) Ositettujen perusmuotoisten, osittamattomien perusmuotoisten ja karsittujen kyselyjen saanti-tarkkuuskäyrät normaalilla relevanssitasolla suomenkielisessä aineistossa.



KUVIO 11. (TUTK) Ositettujen perusmuotoisten, osittamattomien perusmuotoisten ja karsittujen kyselyjen saanti-tarkkuuskäyrät tiukalla relevanssitasolla suomenkielisessä aineistossa.

Yhteenvetona voidaan todeta, että englanninkielisessä aineistossa menetelmät olivat tuloksellisuudeltaan hyvin samankaltaiset. Suomenkielisessä aineistossa osittava perusmuotoistaminen ja perusmuo-

toistaminen olivat tuloksellisuudeltaan hyvin samankaltaiset. Sen sijaan osittava perusmuotoistaminen ja karsinta sekä perusmuotoistaminen ja karsinta poikkesivat kahdella relevanssitasolla tuloksellisuudeltaan toisistaan jopa niin paljon, että niiden väliltä löytyi käytännössä havaittavat erot osittavan perusmuotoistamisen ja perusmuotoistamisen eduksi. Sen lisäksi karsinta oli tuloksellisuudeltaan selvästi muita menetelmiä huonompi.

7.2 Päällekkäisyys

7.2.1 Päällekkäisyys englanninkielisessä aineistossa vertailtaessa kokonaisia tulosjoukkoja

Englanninkielisessä aineistossa perusmuotoistamisella ja karsinnalla saatujen kokonaisten tulosjoukkojen välinen päällekkäisyys on melko suurta, sillä se vaihtelee 8 vertailupisteessä 70 prosentista 74 prosenttiin (ks. Taulukko 8 ja Kuvio 12). Toisin sanoen perusmuotoisen ja karsitun kyselysarjan tulosjoukkojen dokumenteista 70–74 prosenttia on molemmille kyselysarjoille yhteisiä dokumentteja.

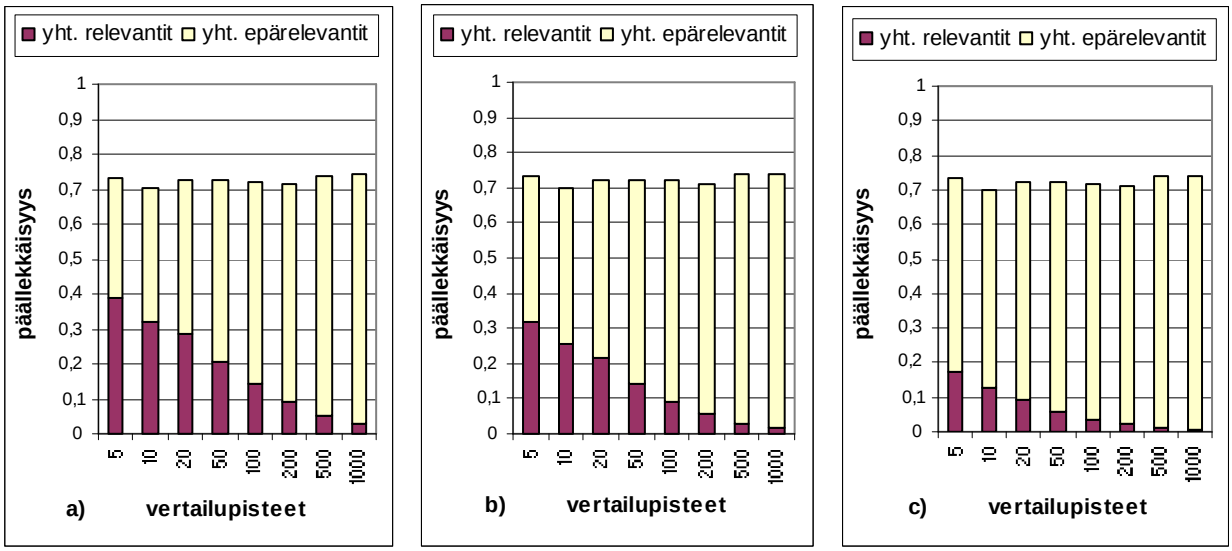
TAULUKKO 8. (TREC) Kokonaisten tulosjoukkojen pareittaisissa vertailuissa yhteisiksi osoittautuneiden dokumenttien määrä eri vertailupisteissä.

ver-tailu-pisteet	perusmuotoinen vs. karsittu
5	0,73
10	0,7
20	0,72
50	0,72
100	0,72
200	0,71
500	0,74
1000	0,74

Tulosjoukoille yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin englanninkielisessä aineistossa

Äskeistä tarkastelua päällekkäisyyden määrästä syvennetään Kuvion 12 avulla, joka havainnollistaa sitä, miten suuri osuus kokonaisille tulosjoukoille yhteisistä dokumenteista on relevantteja ja vastaavasti epärelevantteja.

Kun katsotaan englanninkielisen aineiston osalta, miten suuri osuus verrattaville tulosjoukoille yhteisistä dokumenteista on relevantteja ja miten suuri osuus epärelevantteja dokumentteja, nähdään, että relevantteja dokumentteja on tulosjoukoissa eniten liberaalilla relevanssitasolla (ks. Kuvio 12). Liberaalilla relevanssitasolla relevanttien määrä vaihtelee 39 prosentista 3 prosenttiin, normaalilla relevanssitasolla 32 prosentista 2 prosenttiin ja tiukalla relevanssitasolla 18 prosentista 1 prosenttiin. Relevanttien dokumenttien määrä siis laskee relevanssitasoittain siten, että eniten relevantteja on liberaalilla ja vähiten tiukalla relevanssitasolla. Tämä selittyy saantikannan pienenemisellä relevanssitasoittain. TRECin saantikanta sisältää nimittäin liberaalilla relevanssitasolla 2403, normaalilla relevanssitasolla 1206 ja tiukalla relevanssitasolla 394 relevanttia dokumenttia. Epärelevanttien dokumenttien määrä käyttäytyy luonnollisesti päinvastaisesti, sillä kokonaisille tulosjoukoille yhteisistä dokumenteista epärelevantteja dokumentteja on liberaalilla relevanssitasolla 34–71 prosenttia, normaalilla relevanssitasolla 41–73 prosenttia ja tiukalla relevanssitasolla 56–74. Liberaalilla relevanssitasolla relevanttien dokumenttien osuus verrattaville tulosjoukoille yhteisistä dokumenteista on suurempi kuin epärelevanttien osuus vain yhdessä vertailupisteessä eli vertailupisteessä 5. Normaalin ja tiukan relevanssitason vertailupisteissä epärelevantteja on aina enemmän kuin relevantteja. Niinpä voidaan todeta, että vaikka englanninkielisen aineiston tarkkuuskeskiarvot olivat hieman yllättäviä, jakautuivat tulosjoukot relevantteihin ja epärelevantteihin dokumentteihin odotetunlaisesti.



KUVIO 12. (TREC) Perusmuotoisten ja karsittujen kyselysarjojen tulosjoukoille yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin a) liberaalilla, b) normaalilla ja c) tiukalla relevanssitasolla.

7.2.2 Päällekkäisyys suomenkielisessä aineistossa vertailtaessa kokonaisia tulosjoukkoja

Ensin tarkastellaan millä menetelmillä saatujen tulosjoukkojen välinen päällekkäisyys on suurinta, toiseksi suurinta ja vähäisintä, kun käytössä on suomenkielinen aineisto ja keskenään vertaillaan kokonaisia tulosjoukkoja. Eniten päällekkäisyyttä on ositetun perusmuotoisen ja perusmuotoisen kyselysarjan välillä (87–94 %). Toiseksi eniten päällekkäisyyttä on ilman ositusta tapahtuvan perusmuotoistamisen ja karsinnan välillä (53–61 %). Vähäisintä päällekkäisyys on ositetun perusmuotoisen ja karsitun kyselysarjan välillä (47–57 %) (ks. Taulukko 9 ja Kuvio 13). Päällekkäisyys vaihteli kunkin pareittaisen vertailun kohdalla korkeintaan 10 prosenttiyksikön verran, kun päällekkäisyyttä tarkasteltiin yli kaikkien 8 vertailupisteen.

Vertailtavista kyselysarjoista samankaltaisimmat kyselyt ja samankaltaisimmat kyselyiden hakuavaimet löytyvät ositetusta perusmuotoisesta ja perusmuotoisesta kyselysarjasta. Kyselysarjojen suuresta samankaltaisuudesta johtuen ei ole yllättävää, että päällekkäisyys on suurinta juuri näiden kyselysarjojen välillä (ks. Taulukko 9 ja Kuvio 13). Kun kyseisiä kyselysarjoja vertaillaan keskenään, nähdään, mikä on pelkän osituksen vaikutus päällekkäisyyteen. Suuresta päällekkäisyydestä huolimatta osittamisella näyttää olevan vaikutusta, sillä tulosjoukot eivät mene täysin päällekkäin edes näitä kyselysarjoja vertailtaessa. Päällekkäisyys näiden kyselysarjojen välillä vaihtelee nimittäin 87

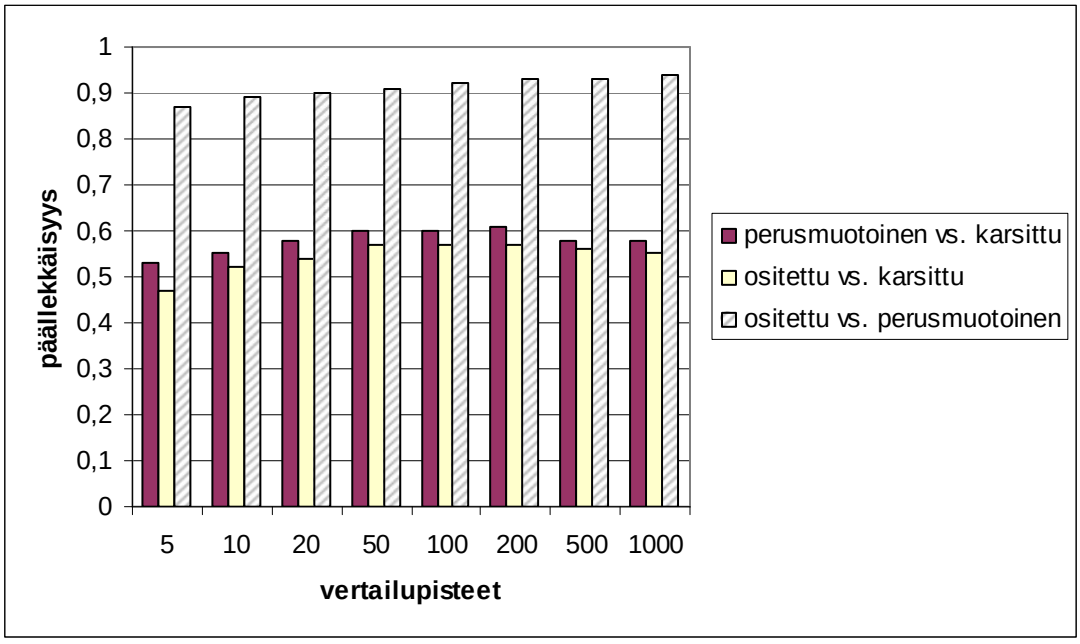
prosentista 94 prosenttiin. Toisin sanoen ositetun perusmuotoisen ja perusmuotoisen kyselysarjan tulosjoukkojen dokumenteista 87–94 prosenttia on kummallekin tulosjoukolle yhteisiä dokumentteja.

Osittamisen vaikutus päällekkäisyyteen nähdään myös, kun verrataan ositetun perusmuotoisen ja karsitun kyselysarjan tulosjoukkojen päällekkäisyyttä osittamattoman perusmuotoisen ja karsitun kyselysarjan tulosjoukkojen päällekkäisyyteen. Ositetun perusmuotoisen ja karsitun kyselysarjan välinen päällekkäisyys vaihtelee 47 prosentista 57:ään prosenttiin, jääden kaikissa vertailupisteissä alle 60 prosentin (ks. Taulukko 9 ja Kuvio 13). Vastaava osittamattoman perusmuotoisen ja karsitun kyselysarjan välinen päällekkäisyys vaihtelee 53:sta prosentista 61:een prosenttiin. Kun ositetun perusmuotoisen ja karsitun kyselysarjan välistä päällekkäisyyttä verrataan osittamattoman perusmuotoisen ja karsitun kyselysarjan väliseen päällekkäisyyteen jokaisessa vertailupisteessä, jotta nähtäisiin osituksen vaikutus päällekkäisyyteen, nähdään osituksen vaikutuksen vaihtelevan 2 prosenttiyksiköstä 6 prosenttiyksikköön (ks. Taulukko 9).

Eroja kyselysarjojen tulosjoukkojen päällekkäisyyksissä selittävät erot kyselysarjojen tuloksellisuudessa. Vaikka osittava perusmuotoistaminen ja perusmuotoistaminen ilman ositusta ovat menetelminä samankaltaisia ja ne olivat myös tuloksellisuudeltaan hyvin samankaltaisia, eivät ne olleet tuloksellisuudeltaan täysin identtisiä, sillä ne eivät löytäneet täysin samoja dokumentteja. Tästä johtuen eivät myöskään niillä saadut tulosjoukot voineet mennä täysin päällekkäin.

TAULUKKO 9. (TUTK) Kokonaisten tulosjoukkojen pareittaisissa vertailuissa yhteisiksi osoittautuneiden dokumenttien määrä eri vertailupisteissä.

vertailupisteet	perusmuotoinen vs. karsittu	ositettu perusmuotoinen vs. karsittu	kahden ensin mainitun pareittaisen vertailun ero %-yksiköissä	ositettu perusmuotoinen vs. perusmuotoinen
5	0,53	0,47	6	0,87
10	0,55	0,52	3	0,89
20	0,58	0,54	4	0,9
50	0,6	0,57	3	0,91
100	0,6	0,57	3	0,92
200	0,61	0,57	4	0,93
500	0,58	0,56	2	0,93
1000	0,58	0,55	3	0,94



KUVIO 13. (TUTK) Kokonaisten tulosjoukkojen pareittaisissa vertailuissa tulosjoukoille yhteisiksi osoittautuneiden dokumenttien määrä.

Lopuksi päällekkäisyyden ja tuloksellisuuden suhdetta tarkastellaan vielä lisää. Tarkastelu tehdään aiemmin, luvussa 7.1.2, esitettyjen tarkkuuskeskiarvojen avulla. Tulosjoukkojen päällekkäisyys näyttää sen perusteella olevan suurinta niiden menetelmien välillä, joiden tuloksellisuudet erosivat vähiten toisistaan ja pienintä niiden menetelmien välillä, joiden tuloksellisuudessa oli eniten eroa (ks. Taulukko 7 ja Kuvio 13). Päällekkäisyyshän oli suurinta ja tuloksellisuuserot pienimpiä ositetun perusmuotoisen ja perusmuotoisen kyselysarjan välillä. Päällekkäisyys oli vastaavasti pienintä ja tuloksellisuuserot suurimpia ositetun perusmuotoisen ja karsitun kyselysarjan välillä.

Tulosjoukoille yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin suomenkielisessä aineistossa

Kun katsotaan suomenkielisen aineiston osalta, miten suuri osuus kokonaisille tulosjoukoille yhteisistä dokumenteista on relevantteja ja miten suuri osa epärelevantteja dokumentteja, nähdään, että relevantteja löytyy eniten ositetun perusmuotoisen ja perusmuotoisen kyselysarjan tulosjoukoista liberaalilla relevanssitasolla (ks. Kuviot 14, 15 ja 16). Relevantteja dokumentteja on vähemmän perusmuotoisen ja karsitun ja vielä vähemmän ositetun perusmuotoisen ja karsitun kyselysarjan tulosjoukkojen yhteisiksi osoittautuneista dokumenteista. Relevanttien dokumenttien määrä laskee relevanssitasoittain siten, että relevanttien dokumenttien määrä on suurinta aina liberaalilla relevanssitasolla, toiseksi suurinta normaalilla relevanssitasolla ja vähäisintä tiukalla relevanssitasolla (ks. Kuviot 14, 15 ja 16). Määrän lasku relevanssitasoittain selittyy tietokannan saantikannan ominaisuuksilla, sillä saantikanta pienenee relevanssitasoittain. TUTKissa saantikanta sisältää liberaalilla relevanssitasolla 1953, normaalilla relevanssitasolla 1066 ja tiukalla relevanssitasolla 366 relevanttia dokumenttia. Saantikannan koon pienentyessä tulee kyselyjen täsmäytyminen relevantteihin dokumentteihin aina vain vaikeammaksi.

Ositetun perusmuotoisen ja perusmuotoisen kyselysarjan pareittaisessa vertailussa relevanttien määrä tulosjoukoille yhteisistä dokumenteista vaihtelee 62 prosentista 1 prosenttiin, kun asiaa tarkastellaan yli kaikkien relevanssitasojen. Perusmuotoisen ja karsitun kyselysarjan pareittaisessa vertailussa relevanttien määrä vaihtelee 37 prosentista 1 prosenttiin, kun asiaa tarkastellaan yli kaikkien relevanssitasojen. Ositetun perusmuotoisen ja karsitun kyselysarjan pareittaisessa vertailussa relevanttien määrä vaihtelee 35 prosentista 1 prosenttiin, kun asiaa tarkastellaan yli kaikkien relevanssitasojen. Tarkem-

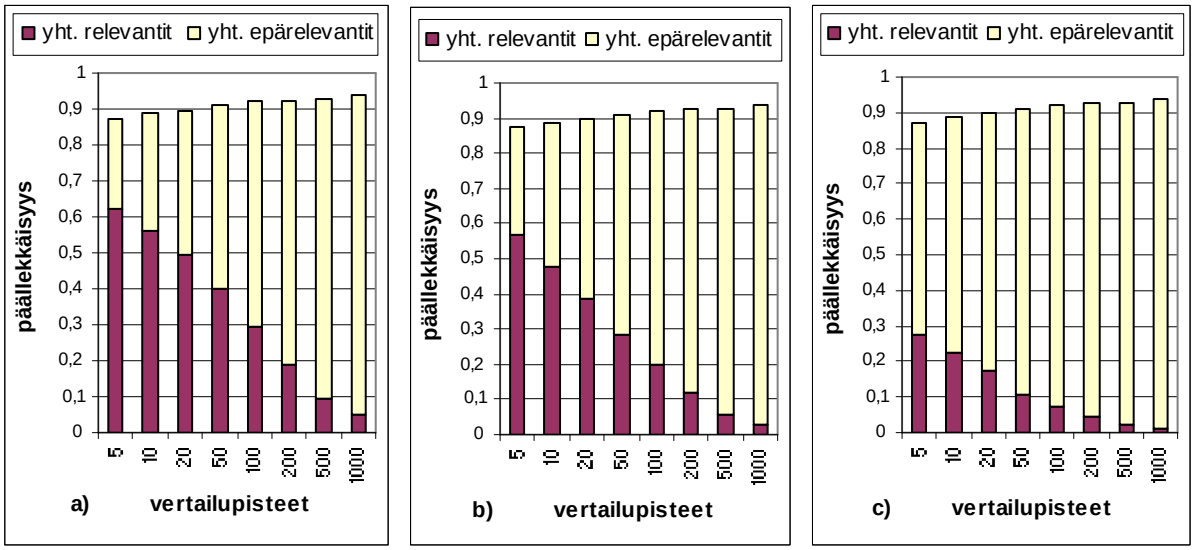
mat tiedot jokaisesta relevanssitasosta esitetään Taulukossa 10. Epärelevanttien dokumenttien määrä noudattelee luonnollisesti päinvastaista järjestystä. Ositetun perusmuotoisen ja perusmuotoisen kyselysarjan pareittaisessa vertailussa epärelevanttien määrä tulosjoukoille yhteisistä dokumenteista vaihtelee 25 prosentista 93 prosenttiin, kun asiaa tarkastellaan yli kaikkien relevanssitasojen. Perusmuotoisen ja karsitun kyselysarjan pareittaisessa vertailussa epärelevanttien määrä vaihtelee 15 prosentista 57 prosenttiin, kun asiaa tarkastellaan yli kaikkien relevanssitasojen. Ositetun perusmuotoisen ja karsitun kyselysarjan pareittaisessa vertailussa epärelevanttien määrä vaihtelee 12 prosentista 54 prosenttiin, kun asiaa tarkastellaan yli kaikkien relevanssitasojen.

TAULUKKO 10. Tulosjoukoille yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin dokumentteihin relevanssitasoittain a) ositettujen perusmuotoisten ja perusmuotoisten kyselysarjojen b) perusmuotoisten ja karsittujen kyselysarjojen sekä c) ositettujen perusmuotoisten ja karsittujen kyselysarjojen pareittaisissa vertailuissa.

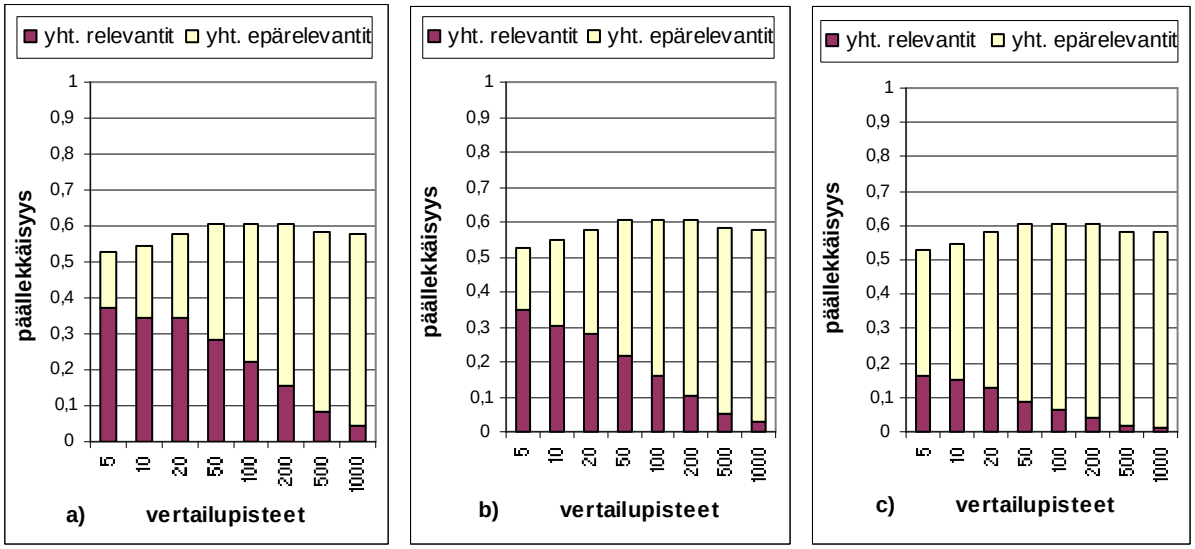
A)	**Vaihteluväli vertailupisteissä**	
	Relevantit	Epärelevantit
Liberaali	5–62 %	25–89%
Normaali	3–57 %	31–91%
Tiukka	1–27 %	60–93%
B)	**Vaihteluväli vertailupisteissä**	
	Relevantit	Epärelevantit
Liberaali	5–37 %	15–53 %
Normaali	3–35 %	18–55 %
Tiukka	1–16 %	37–57 %
C)	**Vaihteluväli vertailupisteissä**	
	Relevantit	Epärelevantit
Liberaali	4–35 %	12–51 %
Normaali	3–33 %	15–53 %
Tiukka	1–15 %	32–54 %

Relevanttien dokumenttien määrän pieneneminen on nähtävissä myös siten, että liberaalilla relevanssitasolla relevanttien dokumenttien osuus kokonaisten tulosjoukkojen yhteisistä dokumenteista on suurempi kuin epärelevanttien osuus vain vertailupisteissä 5, 10 ja 20 tarkasteltiinpa mitä tahansa kyselysarjojen pareittaista vertailua (ks. Kuviot 14, 15 ja 16). Normaalilla relevanssitasolla relevantteja dokumentteja on verrattaville tulosjoukoille yhteisistä dokumenteista enemmän kuin epärelevant-

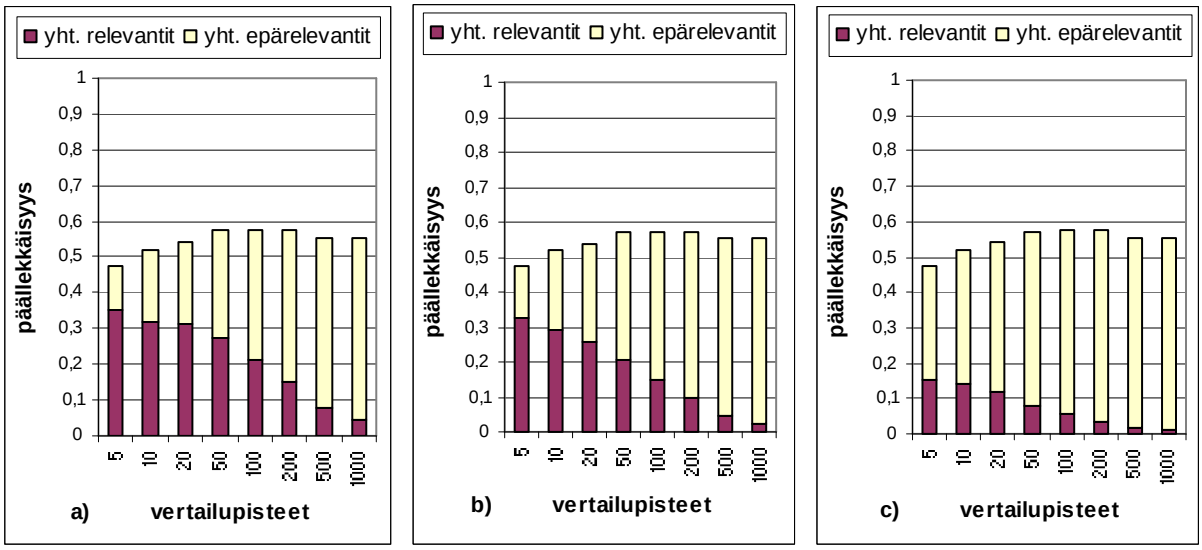
teja vain vertailupisteissä 5 ja 10 katsottiinpa mitä tahansa kyselysarjojen pareittaista vertailua. Tiukalla relevanssitasolla on epärelevantteja dokumentteja aina enemmän kuin relevantteja.



KUVIO 14. (TUTK) Ositettujen perusmuotoisten ja perusmuotoisten kyselysarjojen tulosjoukkojen vertailussa yhteisiksi osoittautuneiden dokumenttien jakautuminen relevantteihin ja epärelevantteihin a) liberaalilla, b) normaalilla ja c) tiukalla relevanssitasolla.



KUVIO 15. (TUTK) Perusmuotoisten ja karsittujen kyselysarjojen tulosjoukkojen vertailussa yhteisiksi osoittautuneiden dokumenttien jakautuminen relevantteihin ja epärelevantteihin a) liberaalilla, b) normaalilla ja c) tiukalla relevanssitasolla.



KUVIO 16. (TUTK) Ositettujen perusmuotoisten ja karsittujen kyselysarjojen tulosjoukkojen vertailussa yhteisiksi osoittautuneiden dokumenttien jakautuminen relevantteihin ja epärelevantteihin a) liberaalilla, b) normaalilla ja c) tiukalla relevanssitasolla.

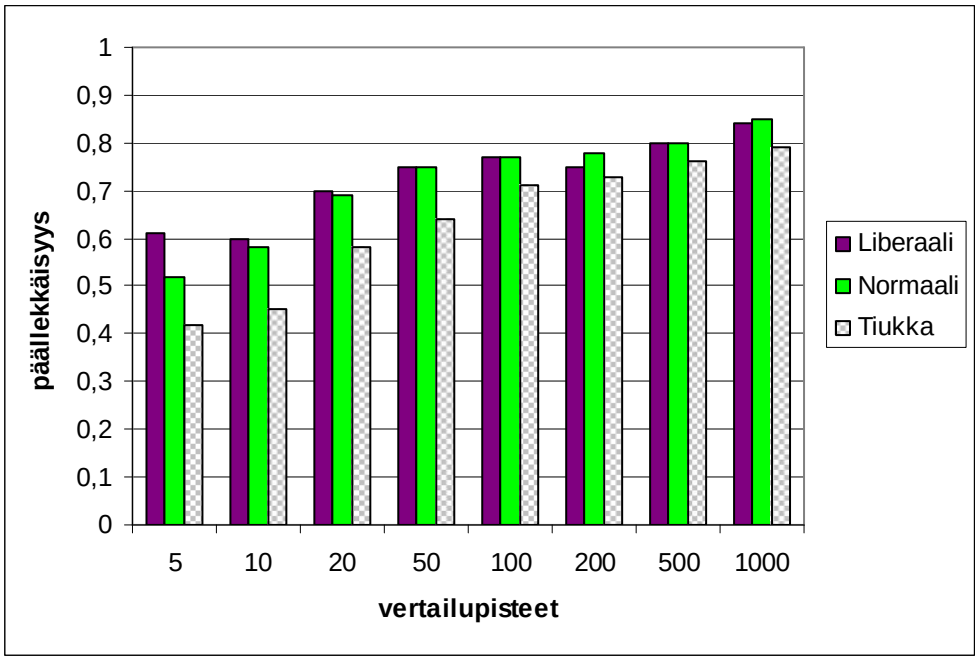
7.2.3 Päällekkäisyys vertailtaessa tulosjoukkojen relevantteja osia

Tulosjoukkojen relevanttien osien päällekkäisyydellä tarkoitetaan kahdelle verrattavalle tulosjoukolle yhteisiä relevantteja dokumentteja, kun vertailukohtana on näiden kahden tulosjoukon kaikki relevantit dokumentit tietyn vertailupisteen rajoissa.

Kun tarkastellaan tulosjoukkojen relevanttien osien päällekkäisyyttä kussakin vertailupisteessä, jaetaan verrattavien tulosjoukkojen yhteisten relevanttien dokumenttien määrä verrattavien tulosjoukkojen kaikkien relevanttien dokumenttien määrällä tai kyseisellä vertailupisteellä, mikäli kaikkien relevanttien määrä on suurempi kuin kyseessä oleva vertailupiste.

Yhteisten relevanttien osuus relevanteista dokumenteista englanninkielisessä aineistossa

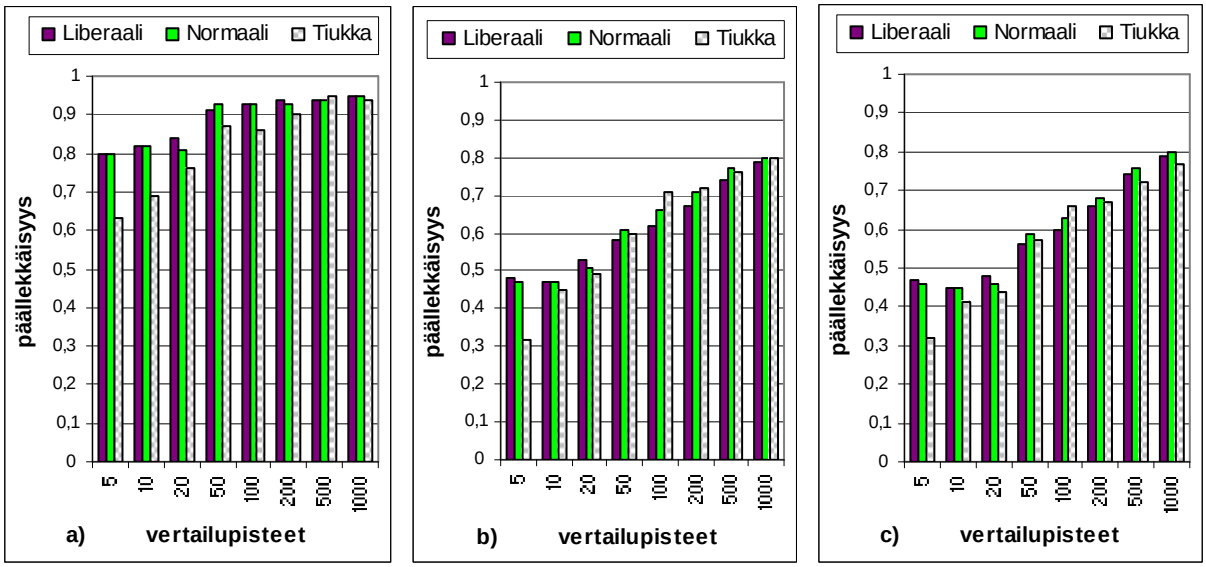
Englanninkielisessä aineistossa karsinnalla ja perusmuotoistamisella saaduille tulosjoukoille yhteisiä relevantteja dokumentteja on 60–84 prosenttia kaikista relevanteista, kun asiaa tarkastellaan liberaalilla relevanssitasolla (ks. Kuvio 17). Normaalilla relevanssitasolla yhteisiä relevantteja on 52–85 prosenttia kaikista relevanteista ja tiukalla relevanssitasolla 42–79 prosenttia kaikista relevanteista.



KUVIO 17. (TREC) Perusmuotoisten ja karsittujen kyselysarjojen yhteisten relevanttien dokumenttien osuus tulosjoukkojen kaikista relevanteista dokumenteista vertailupisteittäin esitettynä.

Yhteisten relevanttien osuus relevanteista dokumenteista suomenkielisessä aineistossa

Ositettujen perusmuotoisten ja perusmuotoisten kyselyjen tulosjoukkojen kaikista relevanteista dokumenteista osoittautuu yhteisiksi relevanteiksi 63–95 prosenttia, kun asiaa tarkastellaan yli kaikkien relevanssitasojen (ks. kuvio 18 a). Tarkemmat tiedot jokaisesta relevanssitasosta esitetään Taulukossa 11.



KUVIO 18. (TUTK) Tulosjoukkojen yhteisten relevanttien dokumenttien osuus kaikista relevanteista vertailupisteittäin, kun toisiinsa verrataan a) ositettuja perusmuotoisia ja perusmuotoisia kyselysarjoja b) osittamattomia perusmuotoisia ja karsittuja kyselysarjoja sekä c) ositettuja perusmuotoisia ja karsittuja kyselysarjoja.

Vertailtaessa toisiinsa osittamattomien perusmuotoisten ja karsittujen kyselysarjojen tulosjoukkojen kaikkia relevantteja dokumentteja nähdään, että niistä yhteisiä relevantteja dokumentteja on 32–80 prosenttia, kun asiaa tarkastellaan yli kaikkien relevanssitasojen (ks. Kuvio 18 b). Tämän pareittaisen vertailun kaikkien relevanssitasojen tulokset esitetään Taulukossa 11.

Osittavalla perusmuotoistamisella ja karsinnalla saatujen tulosjoukkojen kaikista relevanteista dokumenteista 32–80 prosenttia on tulosjoukoille yhteisiä relevantteja dokumentteja, kun asiaa tarkastellaan yli kaikkien relevanssitasojen (ks. Kuvio 18 c). Katso myös Taulukosta 11 kaikkien relevanssitasojen tulokset.

TAULUKKO 11. Tulosjoukkojen kaikista relevanteista dokumenteista yhteisiksi relevanteiksi osoittautuvien osuus verrattaessa toisiinsa a) ositetun perusmuotoisen ja perusmuotoisen, b) perusmuotoisen ja karsitun sekä c) ositetun perusmuotoisen ja karsitun kyselysarjan tulosjoukkoja.

A)	**Vaihteluväli vertailupisteissä**
	Relevantit
Liberaali	80–95 %
Normaali	80–95 %
Tiukka	63–95 %
B)	**Vaihteluväli vertailupisteissä**
	Relevantit
Liberaali	47–79 %
Normaali	47–80 %
Tiukka	32–80 %
C)	**Vaihteluväli vertailupisteissä**
	Relevantit
Liberaali	45–79 %
Normaali	45–80 %
Tiukka	32–77 %

Vaikka kahdessa viimeksi mainitussa pareittaisessa vertailussa relevanttien dokumenttien päällekkäisyyden alimmat ja ylimmät prosenttiosuudet ovat samat, on relevanttien dokumenttien päällekkäisyys kokonaisuudessaan suurempaa osittamattoman perusmuotoisen ja karsitun kyselysarjan pareittaisessa vertailussa (ks. Kuvio 18). Pareittaiset vertailut asettuvat siis samaan järjestykseen verrattiinpa sitten toisiinsa kokonaisten tulosjoukkojen välistä päällekkäisyyttä tai tulosjoukkojen relevanttien osien päällekkäisyyttä, sillä suurinta kummassakin tapauksessa on ositettujen perusmuotoisten ja perusmuotoisten kyselysarjojen tulosjoukkojen päällekkäisyys, toiseksi suurinta perusmuotoisten ja karsittujen kyselysarjojen tulosjoukkojen päällekkäisyys ja vähäisintä ositettujen perusmuotoisten ja karsittujen kyselysarjojen tulosjoukkojen päällekkäisyys (ks. Kuvio 18 ja Kuvio 13).

7.3 Päällekkäisyyden taustalla olevien tekijöiden selvittäminen

Tavallisesti yhteisten relevanttien dokumenttien määrää laskee relevanssitasoittain. Tämä selittyy sillä että tietokannan saantikanta pienenee relevanssitasoittain. Saantikannan pienentyessä relevanttien dokumenttien saaminen tulosjoukkoihin tulee aina vain vaikeammaksi, jolloin myös se on harvinaisempaa, että nuo tulosjoukkoihin poimitut relevantit dokumentit olisivat keskenään päällekkäisiä. Saantikannan koko vaikuttaa siis tuloksellisuuteen ja sitä kautta päällekkäisyyteen. Näin ollen tiukalla relevanssitasolla pitäisi olla vähiten yhteisiä relevantteja dokumentteja ja vastaavasti liberaalilla relevanssitasolla pitäisi yhteisten relevanttien määrän olla suurin. Koska tuloksellisuus ja päällekkäisyys liittyvät toisiinsa, kannattaa tuloksellisuutta ja saanti-tarkkuusarvoja tarkastella uudelleen, aiempaa tässä tulosluvussa tehtyä tarkastelua tarkemmin.

7.3.1 Suomenkielisten kyselysarjojen saanti-tarkkuusarvojen lähempi tarkastelu

Kyselysarjojen saanti-tarkkuusarvoja katsotaan lähemmin, koska ne voivat sisältää tietoa, joka ei käynyt ilmi pelkistä tarkkuuden keskiarvoista, joita tarkasteltiin tässä seitsemännessä luvussa jo aiemmin. Tässä keskitytään etenkin suomenkielisten kyselysarjojen saanti-tarkkuusarvojen tarkasteluun. Englanninkielisten kyselysarjojen saanti-tarkkuusarvot löytyvät kokonaisuudessaan liitteistä. Taulukoissa 12, 13 ja 14 verrataan suomenkielisessä aineistossa käytettyjen menetelmien saanti-tarkkuusarvoja toisiinsa ja katsotaan, miten paljon eri menetelmät eroavat toisistaan. Aluksi tarkastellaan saanti-tarkkuusarvojen eroja kussakin pareittaisessa vertailussa. Lopulta vertaillaan pareittaisia vertailuja toisiinsa. Englanninkielisessä aineistossa tehtiin vain yksi pareittainen vertailu, joten tehtyä pareittaista vertailua ei voida verrata toisen pareittaisen vertailun kanssa. Näin ollen englanninkielisen aineiston saanti-tarkkuusarvoja ei tarkastella tässä luvussa sen enempää.

Ensin lasketaan paljonko tarkkuudet eroavat ositetun perusmuotoisen ja perusmuotoisen kyselysarjan välillä. Toisekseen tarkastellaan eroja osittamattoman perusmuotoisen ja karsitun kyselysarjan tarkkuuksissa. Lisäksi lasketaan, miten paljon ositetun perusmuotoisen ja karsitun kyselysarjan tarkkuudet eroavat toisistaan. Sen jälkeen katsotaan, millä relevanssitasolla havaitut erot ovat suurimpia ja millä pienimpiä. Suurin ero on ilmaistu Taulukoissa 12, 13 ja 14 lihavoinnin ja pienin ero alleviivauksen avulla. Keskeisintä on katsoa, ovatko erot keskittyneet johdonmukaisesti jollekin relevanssita-

solle. Jos näin on, millä relevanssitasolla kahden menetelmän väliset erot ovat useimmiten suurimmat ja useimmiten pienimmät.

Ensin tarkastellaan eroja ositetun perusmuotoisen ja perusmuotoisen kyselysarjan välillä. Erot ositetun perusmuotoisen ja perusmuotoisen kyselysarjan tarkkuudessa ovat pienimmät liberaalilla relevanssitasolla, jossa ero on pienin neljällä saantitasolla (ks. Taulukko 12). Erot ositetun perusmuotoisen ja perusmuotoisen tarkkuudessa ovat suurimmat viidellä normaalin relevanssitason saantitasolla.

TAULUKKO 12. (TUTK) Ositetun perusmuotoisen ja perusmuotoisen kyselysarjan saantitarkkuusarvot liberaalilla, normaalilla ja tiukalla relevanssitasolla.

	Ositettu perusmuotoinen kyselysarja				**Perusmuotoinen kyselysarja**			**Relevanssitasojen väliset erot**		
Saantitaso	Liberaali	Normaali	Tiukka		Liberaali	Normaali	Tiukka	Ero liberaalilla relevanssitasolla	Ero normaalilla relevanssitasolla	Ero tiukalla relevanssitasolla
10	69,8	69,8	50,8		70,2	66,2	48,2	-0,4	**3,6**	2,6
20	62,2	59,8	40,2		62,2	56,9	38,7	0	2,9	**3,8**
30	56,1	52,1	33,7		57,4	51,5	33	**-1,3**	0,6	0,7
40	48,2	44,7	29,7		48,4	45,4	30,3	-0,2	**-0,7**	-0,6
50	40	38,3	28,4		39	38,1	29	**1**	0,2	-0,6
60	31,9	29,5	20,6		31,7	30,2	21,2	0,2	**-0,7**	-0,6
70	24,6	22,9	15,5		24,8	22,6	15,7	-0,2	**0,3**	-0,2
80	17,9	17,1	12,7		17,5	16,6	12,7	0,4	**0,5**	0
90	11,6	10,3	7,8		11	9,7	8	0,6	0,6	-0,2
100	2,5	5,2	6,1		2,4	5,3	6,4	0,1	-0,1	**-0,3**

Toiseksi tarkastellaan eroja perusmuotoisen ja karsitun kyselysarjan välillä. Perusmuotoisen ja karsitun kyselysarjan välillä erot tarkkuudessa ovat pienimmät tiukalla relevanssitasolla, jossa ero on pienin yhdeksällä saantitasolla kymmenestä (ks. Taulukko 13). Erot perusmuotoisen ja karsitun kyselysarjan tarkkuudessa ovat suurimmat seitsemällä liberaalin relevanssitason saantitasolla ja kolmella normaalin relevanssitason saantitasolla.

TAULUKKO 13. (TUTK) Perusmuotoisen ja karsitun kyselysarjan saanti-tarkkuusarvot liberaalilla, normaalilla ja tiukalla relevanssitasolla.

	Perusmuotoinen kyselysarja				**Karsittu kyselysarja**			**Relevanssitasojen väliset erot**		
Saantitaso	Liberaali	Normaali	Tiukka		Liberaali	Normaali	Tiukka	Ero liberaalilla relevanssitasolla	Ero normaalilla relevanssitasolla	Ero tiukalla relevanssitasolla
10	70,2	66,2	48,2		60,5	59,5	41,6	**9,7**	6,7	6,6
20	62,2	56,9	38,7		53,1	49,9	36,4	**9,1**	7	2,3
30	57,4	51,5	33		44,3	42,7	32,7	**13,1**	8,8	0,3
40	48,4	45,4	30,3		35,5	32,7	25,3	**12,9**	12,7	5
50	39	38,1	29		27,9	26	22,9	11,1	**12,1**	6,1
60	31,7	30,2	21,2		22,3	22,5	14,9	**9,4**	7,7	6,3
70	24,8	22,6	15,7		17,3	15,7	10,2	**7,5**	6,9	5,5
80	17,5	16,6	12,7		13,5	12,1	9,1	4	**4,5**	3,6
90	11	9,7	8		8	7,5	6,4	**3**	2,2	1,6
100	2,4	5,3	6,4		2,3	2,9	4,1	0,1	**2,4**	2,3

Kolmanneksi tarkastellaan eroja ositetun perusmuotoisen ja karsitun kyselysarjan välillä. Ositetun perusmuotoisen ja karsitun kyselysarjan välillä erot tarkkuudessa ovat pienimmät tiukalla relevanssitasolla, jossa ero on pienin yhdeksällä saantitasolla kymmenestä (ks. Taulukko 14). Erot ositetun perusmuotoisen ja karsitun kyselysarjan tarkkuudessa ovat suurimmat viidellä liberaalin relevanssitason saantitasolla ja viidellä normaalin relevanssitason saantitasolla.

TAULUKKO 14. (TUTK) Ositetun perusmuotoisen ja karsitun kyselysarjan saanti-tarkkuusarvot liberaalilla, normaalilla ja tiukalla relevanssitasolla.

	Ositettu perusmuotoinen kyselysarja				**Karsittu kyselysarja**			**Relevanssitasojen väliset erot**		
Saantitaso	Liberaali	Normaali	Tiukka		Liberaali	Normaali	Tiukka	Ero liberaalilla relevanssitasolla	Ero normaalilla relevanssitasolla	Ero tiukalla relevanssitasolla
10	69,8	69,8	50,8		60,5	59,5	41,6	9,3	**10,3**	9,2
20	62,2	59,8	40,2		53,1	49,9	36,4	9,1	**9,9**	3,8
30	56,1	52,1	33,7		44,3	42,7	32,7	**11,8**	9,4	1
40	48,2	44,7	29,7		35,5	32,7	25,3	**12,7**	12	4,4
50	40	38,3	28,4		27,9	26	22,9	12,1	**12,3**	5,5
60	31,9	29,5	20,6		22,3	22,5	14,9	**9,6**	7	5,7
70	24,6	22,9	15,5		17,3	15,7	10,2	**7,3**	7,2	5,3
80	17,9	17,1	12,7		13,5	12,1	9,1	4,4	**5**	3,6
90	11,6	10,3	7,8		8	7,5	6,4	**3,6**	2,8	1,4
100	2,5	5,2	6,1		2,3	2,9	4,1	0,2	**2,3**	2

Erot tarkkuuksissa eivät näyttäisi olevan johdonmukaisesti suurimpia tai pienimpiä jollakin tietyllä relevanssitasolla, ei ainakaan, kun asiaa tarkastellaan kaikkien kolmen pareittaisen vertailun osalta. Kuitenkin kolmesta pareittaisesta vertailusta kaksi näyttäisi käyttäytyvän keskenään samankaltaisesti. Erot perusmuotoisen ja karsitun kyselysarjan sekä ositetun perusmuotoisen ja karsitun kyselysarjan välillä näyttäisivät nimittäin olevan keskenään aika samankaltaiset. Näiden kahden pareittaisen vertailun osalta voidaan sanoa, että saanti-tarkkuusarvojen erojen tarkastelussa erot ovat useimmiten, joskaan eivät aina, suurimmat liberaalilla relevanssitasolla ja pienimmät tiukalla relevanssitasolla. Kolmannen pareittaisen vertailun tarkkuusarvojen suurimmat ja pienimmät erot ovat jakautuneet eri relevanssitasoille paljon hajanaisemmin. Kun tarkkuuksien välisiä eroja tarkastellaan kunkin pareittaisen vertailun osalta, nähdään että erot tarkkuuksissa ovat pienimpiä (-1,3–3,8) ositetun perusmuotoisen ja perusmuotoisen kyselysarjan välillä. Kahdessa muussa pareittaisessa vertailussa tarkkuudet eroavat toisistaan enemmän. Näille kahdelle pareittaiselle vertailulle näyttää olevan yhteistä se, että kummassakin vertailussa on mukana karsittu kyselysarja. Kun vielä muistetaan, että päällekkäisyys ja tuloksellisuus liittyvät toisiinsa, lienee järkevää tarkastella karsitun kyselysarjan tuloksellisuutta näissä kahdessa pareittaisessa vertailussa. Taulukon 13 saanti-tarkkuusarvoista käykin ilmi, että perus-

muotoinen kyselysarja on tuloksellisuudeltaan aina karsittua kyselysarjaa parempi. Samalla tavalla Taulukosta 14 nähdään, että ositettu perusmuotoinen kyselysarja on aina karsittua kyselysarjaa parempi.

Kuten aiemmin luvussa 7.1.2 todettiin, tässä tutkielmassa käytettyjen suomenkielisten kyselyiden keskitarkkuudet laskevat relevanssitasoittain siten, että keskitarkkuus on korkein liberaalilla ja matalin tiukalla relevanssitasolla. Silloin myös potentiaalisten tulosjoukoille yhteisten relevanttien dokumenttien määrä pienenee asteittain. Toisin sanoen tiukan relevanssitason dokumentteja on alun alkaen vaikeampi saada tulosjoukkoon kuin normaalin ja liberaalin relevanssitason dokumentteja. Sen seurauksena tulosjoukkoon tiukalla relevanssitasolla saaduista relevanteista dokumenteista vielä pienempi osuus osoittautuu verrattaville tulosjoukoille yhteisiksi. Tämä vaikuttaa kaikissa pareittaisissa vertailuissa siten, että päällekkäisyys alenee relevanssitasoittain. Tuloksellisuuden ja tulosjoukkojen päällekkäisyyden pienenemiseen vaikuttaa relevanssitasojen lisäksi karsittu kyselysarja. Se että TUTKin karsittu kyselysarja on tässä tutkielmassa tuloksellisuudeltaan selkeästi verrokkejaan heikompi jokaisella relevanssitasolla, tarkoittaa ettei se saa poimittua tulosjoukkoon relevantteja dokumentteja yhtä hyvin kuin muut kyselysarjat. Koska karsitulla kyselysarjalla on vaikeaa saada relevantteja dokumentteja tulosjoukkoon, on niistä vielä pienempi osuus yhteisiä verrattavien tulosjoukkojen kesken ja päällekkäisyys jää pienemmäksi. Päällekkäisyyden määrän lasku pareittaisissa vertailuissa, joissa toisena osapuolena on karsitun kyselysarjan tulosjoukko, johtuu siis karsitun kyselysarjan heikommasta tuloksellisuudesta.

8 Keskustelua

Tässä tutkielmassa oli tarkoitus tutkia tulosjoukkojen päällekkäisyyttä suomen- ja englanninkielisen aineiston avulla. Niinpä karsituilla kyselyillä tietokannan karsitusta hakemistosta saatuja tulosjoukkoja ja perusmuotoisilla kyselyillä perusmuotoisesta hakemistosta saatuja tulosjoukkoja verrattiin keskenään. Suomenkielisessä aineistossa oli lisäksi tarkoituksena selvittää, miten paljon tulosjoukot ovat päällekkäisiä, kun verrataan keskenään ositettujen ja osittamattomien kyselyversioiden tulosjoukkoja. Pyrkimyksenä oli katsoa, vaihteleeko päällekkäisyyden määrä, kun päällekkäisyyttä tarkastellaan a) vertaamalla toisiinsa kokonaisia tulosjoukkoja, b) vertaamalla toisiinsa vain tulosjoukkojen relevantteja osia. Päällekkäisyyttä tarkasteltiin liberaalilla, normaalilla ja tiukalla relevanssitasolla. Sen lisäksi

päällekkäisyyttä tarkasteltiin 8:ssa eri vertailupisteessä. Tutkielmassa tarkasteltiin päällekkäisyyden lisäksi myös kyselyjen tarkkuuksia. Tämä tutkimus tehtiin laboratorio-olosuhteissa, jossa testikokoelma tarjosi vakioidun tutkimusympäristön ja -olosuhteet.

8.1 Tarkkuudet suomen- ja englanninkielisessä aineistossa

Englanninkielisessä aineistossa karsitun kyselysarjan tarkkuuskeskiarvot olivat 1–1,5 prosenttiyksikköä parempia kuin perusmuotoisen kyselysarjan. Niinpä tutkielmassa voitiin todeta, että englanninkielisessä aineistossa perusmuotoistaminen ja karsinta olivat tuloksellisuudeltaan hyvin samankaltaiset menetelmät. Tältä osin tämän tutkielman tulokset olivat linjassa aiempien tutkimusten tulosten kanssa. Englanninkielisen aineiston tarkkuuskeskiarvot olivat jossain määrin yllättäviä, sillä tuloksellisuus ei laskenut johdonmukaisesti relevanssitasoittain. Tavallisuudesta poiketen keskitarkkuus oli korkein normaalilla relevanssitasolla, toiseksi korkein tiukalla relevanssitasolla ja matalin liberaalilla relevanssitasolla. Poikkeukselliset tarkkuuskeskiarvot selittyivät sillä, että relevantit ja erittäin relevantit dokumentit löytyivät tulosjoukosta keskimäärin ennen marginaalisesti relevantteja dokumentteja, jotka sijoittuivat tulosjoukossa pääosin relevanttien ja erittäin relevanttien jälkeen.

Suomenkielisessä aineistossa parhaat tarkkuuskeskiarvot saatiin ositetuilla perusmuotoisilla kyselyillä, toiseksi parhaat perusmuotoisilla kyselyillä ja heikoimmat karsituilla kyselyillä. Ositetun perusmuotoisen ja perusmuotoisen kyselysarjan tarkkuuskeskiarvojen väliset erot olivat hyvin pienet. Sen sijaan kahdessa muussa pareittaisessa vertailussa oli havaittavissa suurempia tarkkuuskeskiarvojen välisiä eroja. Kyselysarjojen väliltä löytyvät tuloksellisuuserot olivat nimittäin suurehkoja verrattiinpa karsitun kyselysarjan kanssa sitten ositettua perusmuotoista tai osittamatonta perusmuotoista kyselysarjaa. Erot ositetun perusmuotoisen ja karsitun kyselysarjan välillä, sekä osittamattoman perusmuotoisen ja karsitun kyselysarjan välillä, olivat käytännössä havaittavia liberaalilla ja normaalilla relevanssitasolla. Karsinta oli siis tuloksellisuudeltaan selvästi muita menetelmiä huonompi. Myös suomenkielisen aineiston osalta tässä tutkielmassa käytettyjen menetelmien tulokset olivat linjassa aiempien tutkimustulosten kanssa.

8.2 Päällekkäisyys

8.2.1 Päällekkäisyys englannin- ja suomenkielisessä aineistossa vertailtaessa kokonaisia tulosjoukkoja

Kun pareittaiset vertailut tehtiin englanninkielisessä aineistossa vertailemalla toisiinsa kokonaisia tulosjoukkoja, oli perusmuotoisen ja karsitun kyselysarjan välinen päällekkäisyys melko suurta, sillä se vaihteli 8 vertailupisteessä 70 prosentista 74 prosenttiin.

Kun pareittaiset vertailut tehtiin suomenkielisessä aineistossa vertailemalla toisiinsa kokonaisia tulosjoukkoja, oli päällekkäisyyttä eniten ositetun perusmuotoisen ja perusmuotoisen kyselysarjan välillä (87–94 %). Toiseksi eniten päällekkäisyyttä havaittiin perusmuotoistamisen ja karsinnan välillä (53–61 %). Vähäisintä päällekkäisyys oli ositetun perusmuotoisen ja karsitun kyselysarjan välillä (47–57 %). Päällekkäisyys vaihteli kunkin pareittaisen vertailun kohdalla korkeintaan 10 prosenttiyksikön verran, kun päällekkäisyyttä tarkasteltiin yli kaikkien 8 vertailupisteen. Ositus vaikutti päällekkäisyyteen siten, että tulosjoukot eivät menneet täysin päällekkäin edes silloin, kun toisiinsa verrattiin tulosjoukkoja jotka oli saatu kyselyillä, joiden hakuavaimet erosivat toisistaan vain yhdyssanojen osituksen osalta.

Tutkielmassa käytettyjen kyselyversioiden hakuavainten lähes täydellisen samankaltaisuuden vuoksi oletettiin, ettei päällekkäisyyden määrä jää tässä tutkimuksessa yhtä matalaksi kuin aiemmissa tutkimuksissa, joissa hakuavainten samankaltaisuus on ollut vähäisempää. Kun Lee (1996, 7–9) tarkasteli, missä määrin erilaisilla laajennetuilla kyselyvektoreilla saadut tulosjoukot olivat päällekkäisiä, vaihteli tulosjoukkojen päällekkäisyys 52 prosentista jopa 99 prosenttiin, kun päällekkäisyyttä tarkasteltiin kokonaisissa tulosjoukoissa. Lee teki pareittaisia vertailuja yhteensä 10 kappaletta. Niistä kolmessa päällekkäisyys oli 52–54 prosenttia. Kolmen muun pareittaisen vertailun päällekkäisyys asettui 64–67 prosenttiin. Yhdessä pareittaisessa vertailussa päällekkäisyys oli 74 %. Kahdessa pareittaisessa vertailussa päällekkäisyys oli 88 %. Yhden pareittaisen vertailun päällekkäisyys oli 99 %. (Lee 1996, 9.) Tässä tutkielmassa saadut päällekkäisyydet asettuvat Leen esittämien lukemien välimaastoon.

Leellä (1996, 9) tulosjoukkojen päällekkäisyys oli suurinta niiden kyselyvektorien välillä, jotka olivat vektoreina samankaltaisimmat. On selvää, että tässä tutkielmassa ositettu perusmuotoinen ja perus-

muotoinen kyselysarja ovat keskenään samankaltaisimmat kyselysarjat, kun kyselysarjojen samankaltaisuutta tarkastellaan sekä suomen- että englanninkielisessä aineistossa. Myös päällekkäisyys on suurinta verrattaessa niillä saatuja kokonaisia tulosjoukkoja toisiinsa (87–94 %). Joten tämän tutkielman tulokset näyttäisivät tukevan Leen (1996) havaintoa. Englanninkielisessä aineistossa kokonaisten tulosjoukkojen päällekkäisyys on 70–74 %, joten perusmuotoinen kyselysarja näyttäisi olevan varsin samankaltainen karsitun kyselysarjan kanssa. Suomenkielisessä aineistossa perusmuotoistamisella ja karsinnalla saatujen kokonaisten tulosjoukkojen päällekkäisyys on sen sijaan vaatimattomampaa (53–61 %), joten kyselyt näyttävät poikkeavan toisistaan englanninkielisiä vastineitaan enemmän. Suomenkielisillä kyselyillä saatujen tulosjoukkojen päällekkäisyys ja oletettu kyselyiden samankaltaisuus vähenee entisestään verrattaessa karsitun kyselysarjan kanssa ositettua perusmuotoista kyselysarjaa, sillä päällekkäisyys on silloin vain 47–57 %. Päällekkäisyys näyttää siis olevan suurinta niiden kyselysarjojen välillä, jotka ovat kyselysarjoina samankaltaisimmat.

Tarkasteltaessa päällekkäisyyden ja tuloksellisuuden välistä suhdetta suomenkielisessä aineistossa havaittiin, että kokonaisten tulosjoukkojen päällekkäisyys oli suurinta niiden menetelmien välillä, joiden tuloksellisuudet erosivat vähiten toisistaan ja pienintä niiden menetelmien välillä, joiden tuloksellisuudessa oli eniten eroa. Tästä syystä kokonaisten tulosjoukkojen päällekkäisyys oli vähäisintä niissä suomenkielisen aineiston pareittaisissa vertailuissa, joissa toisena osapuolena oli karsittu kyselysarja. Karsitun kyselysarjan tuloksellisuus oli nimittäin heikompi kuin niiden kahden muun kyselysarjan tuloksellisuus, johon karsittua kyselysarjaa kulloinkin verrattiin.

8.2.2 Tulosjoukoille yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin englannin- ja suomenkielisessä aineistossa

Englanninkielisessä aineistossa katsottiin, miten suuri osuus toisiinsa verrattavien kokonaisten tulosjoukkojen yhteisistä dokumenteista oli relevantteja ja miten suuri osuus epärelevantteja dokumentteja. Silloin nähtiin, että relevanttien dokumenttien määrä laski relevanssitasoittain siten, että eniten relevantteja oli liberaalilla ja vähiten tiukalla relevanssitasolla. Liberaalilla relevanssitasolla relevanttien määrä vaihteli 39 prosentista 3 prosenttiin. Normaalilla relevanssitasolla määrä vaihteli puolestaan 32 prosentista 2 prosenttiin. Tiukalla relevanssitasolla relevanttien määrä vaihteli 18 prosentista 1 prosenttiin. Epärelevanttien dokumenttien määrä käyttäytyi luonnollisesti päinvastaisesti.

Kun katsottiin suomenkielisen aineiston osalta, miten suuri osuus toisiinsa verrattavien kokonaisten tulosjoukkojen yhteisistä dokumenteista oli relevantteja ja miten suuri osa epärelevantteja dokumentteja nähtiin, että relevantteja dokumentteja löytyi eniten ositetun perusmuotoisen ja perusmuotoisen kyselysarjan tulosjoukoista. Relevantteja dokumentteja oli vähemmän perusmuotoisen ja karsitun ja vielä vähemmän ositetun perusmuotoisen ja karsitun kyselysarjan tulosjoukkojen yhteisiksi osoittautuneista dokumenteista. Relevanttien dokumenttien määrä laski kunkin pareittaisen vertailun kohdalla relevanssitasoittain. Ositetun perusmuotoisen ja perusmuotoisen kyselysarjan pareittaisessa vertailussa relevanttien määrä vaihteli 62 prosentista 1 prosenttiin, kun asiaa tarkasteltiin yli kaikkien relevanssitasojen. Perusmuotoisen ja karsitun kyselysarjan pareittaisessa vertailussa relevanttien määrä vaihteli 37 prosentista 1 prosenttiin, kun asiaa tarkasteltiin yli kaikkien relevanssitasojen. Ositetun perusmuotoisen ja karsitun kyselysarjan pareittaisessa vertailussa relevanttien määrä vaihteli 35 prosentista 1 prosenttiin, kun asiaa tarkasteltiin yli kaikkien relevanssitasojen. Epärelevanttien dokumenttien määrä noudatteli luonnollisesti päinvastaista järjestystä. Äskeiset tulokset ovat nähtävissä kunkin relevanssitason osalta Taulukoista 15, 16, 17 ja 18.

8.3 Päällekkäisyys vertailtaessa tulosjoukkojen relevantteja osia englannin- ja suomenkielisessä aineistossa

Kun englanninkielisessä aineistossa verrattiin toisiinsa karsinnalla ja perusmuotoistamisella saatujen tulosjoukkojen relevantteja osia, osoittautui relevanteista dokumenteista yhteisiksi relevanteiksi 60–84 prosenttia, kun asiaa tarkasteltiin liberaalilla relevanssitasolla. Normaalilla relevanssitasolla yhteisiä relevantteja oli 52–85 prosenttia kaikista relevanteista ja tiukalla relevanssitasolla 42–79 prosenttia kaikista relevanteista.

Suomenkielisessä aineistossa ositettujen perusmuotoisten ja perusmuotoisten kyselyjen tulosjoukkojen relevanteista dokumenteista yhteisiä relevantteja oli 63–95 prosenttia, kun asiaa tarkasteltiin yli kaikkien relevanssitasojen. Vertailtaessa toisiinsa perusmuotoisten ja karsittujen kyselysarjojen tulosjoukkojen relevantteja dokumentteja nähtiin, että niistä yhteisiä relevantteja dokumentteja oli 32–80 prosenttia, kun asiaa tarkasteltiin yli kaikkien relevanssitasojen. Osittavalla perusmuotoistamisella ja karsinnalla saatujen tulosjoukkojen relevanteista dokumenteista 32–80 prosenttia oli tulosjoukoille yhteisiä relevantteja dokumentteja, kun asiaa tarkasteltiin yli kaikkien relevanssitasojen. Tarkemmat tulokset kultakin relevanssitasolta on esitetty Taulukoissa 15, 16, 17 ja 18.

Lopuksi esitetään vielä Taulukot 15, 16, 17 ja 18 joissa on yhteenveto kaikkien pareittaisten vertailujen kaikista edellä esitetyistä päällekkäisyyksistä. Sen lisäksi taulukoissa on esitetty kokonaisille tulosjoukoille yhteisistä dokumenteista relevanteiksi ja epärelevanteiksi dokumenteiksi osoittautuneiden dokumenttien keskiarvot yli vertailupisteiden. Keskiarvot yli vertailupisteiden on laskettu myös tarkasteltaessa yhteisten relevanttien dokumenttien osuutta kaikista relevanteista dokumenteista sekä tarkasteltaessa kokonaisten tulosjoukkojen päällekkäisyyttä.

TAULUKKO 15. (TREC) Perusmuotoisen ja karsitun kyselysarjan päällekkäisyyksien yhteenveto.

Päällekkäisyys vertailtaessa kokonaisia tulosjoukkoja (ks. luku 8.2.1)		Yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin (ks. luku 8.2.2)			Yhteisten relevanttien osuus kaikista relevanteista (ks. luku 8.3)	
Vaihteluväli vertailupisteissä		**Vaihteluväli vertailupisteissä**			**Vaihteluväli vertailupisteissä**	
Kaikki		Relevantit	Epärelevantit		Relevantit	
70–74 %	Liberaali	3–39 %	34–71 %	Liberaali	60–84 %	
	Normaali	2–32 %	41–73 %	Normaali	52–85 %	
	Tiukka	1–18 %	56–74 %	Tiukka	42–79 %	
Keskiarvo yli vertailupisteiden		**Keskiarvot yli vertailupisteiden**			**Keskiarvot yli vertailupisteiden**	
Kaikki		Relevantit	Epärelevantit		Relevantit	
72 %	Liberaali	19 %	53 %	Liberaali	73 %	
	Normaali	14 %	58 %	Normaali	72 %	
	Tiukka	7 %	66 %	Tiukka	64 %	

TAULUKKO 16. (TUTK) Ositetun perusmuotoisen ja perusmuotoisen kyselysarjan päällekkäisyyksien yhteenveto.

Päällekkäisyys vertailtaessa kokonaisia tulosjoukkoja (ks. luku 8.2.1)	Yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin (ks. luku 8.2.2)			Yhteisten relevanttien osuus kaikista relevanteista (ks. luku 8.3)	
Vaihteluväli vertailupisteissä		**Vaihteluväli vertailupisteissä**			**Vaihteluväli vertailupisteissä**
Kaikki		Relevantit	Epärelevantit		Relevantit
87–94%	Liberaali	5–62 %	25–89 %	Liberaali	80–95 %
	Normaali	3–57%	31–91 %	Normaali	80–95 %
	Tiukka	1–27 %	60–93 %	Tiukka	63–95 %
Keskiarvo yli vertailupisteiden		**Keskiarvot yli vertailupisteiden**			**Keskiarvot yli vertailupisteiden**
Kaikki		Relevantit	Epärelevantit		Relevantit
91 %	Liberaali	34 %	57 %	Liberaali	89 %
	Normaali	26 %	65 %	Normaali	89 %
	Tiukka	12 %	79 %	Tiukka	83 %

TAULUKKO 17. (TUTK) Perusmuotoisen ja karsitun kyselysarjan päällekkäisyyksien yhteenveto.

Päällekkäisyys vertailtaessa kokonaisia tulosjoukkoja (ks. luku 8.2.1)	Yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin (ks. luku 8.2.2)			Yhteisten relevanttien osuus kaikista relevanteista (ks. luku 8.3)	
Vaihteluväli vertailupisteissä		**Vaihteluväli vertailupisteissä**			**Vaihteluväli vertailupisteissä**
Kaikki		Relevantit	Epärelevantit		Relevantit
53–61%	Liberaali	5–37 %	15–53 %	Liberaali	47–79 %
	Normaali	3–35%	18–55 %	Normaali	47–80 %
	Tiukka	1–16 %	37–57 %	Tiukka	32–80 %
Keskiarvo yli vertailupisteiden		**Keskiarvot yli vertailupisteiden**			**Keskiarvot yli vertailupisteiden**
Kaikki		Relevantit	Epärelevantit		Relevantit
58 %	Liberaali	23 %	35 %	Liberaali	61 %
	Normaali	18 %	39 %	Normaali	63 %
	Tiukka	8 %	50 %	Tiukka	61 %

TAULUKKO 18. (TUTK) Ositetun perusmuotoisen ja karsitun kyselysarjan päällekkäisyyksien yhteenveto.

Päällekkäisyys vertailtaessa kokonaisia tulosjoukkoja (ks. luku 8.2.1)		Yhteisten dokumenttien jakautuminen relevantteihin ja epärelevantteihin (ks. luku 8.2.2)			Yhteisten relevanttien osuus kaikista relevanteista (ks. luku 8.3)	
Vaihteluväli vertailupisteissä			**Vaihteluväli vertailupisteissä**			**Vaihteluväli vertailupisteissä**
Kaikki			Relevantit	Epärelevantit		Relevantit
47–57%		Liberaali	4–35 %	12–51 %	Liberaali	45–79 %
		Normaali	3–33%	15–53 %	Normaali	45–80 %
		Tiukka	1–15 %	32–54 %	Tiukka	32–77 %
Keskiarvo yli vertailupisteiden			**Keskiarvot yli vertailupisteiden**			**Keskiarvot yli vertailupisteiden**
Kaikki			Relevantit	Epärelevantit		Relevantit
54 %		Liberaali	22 %	33 %	Liberaali	59 %
		Normaali	17 %	37 %	Normaali	60 %
		Tiukka	8 %	47 %	Tiukka	57 %

Keskustelua -luvun tarkoituksena on tutkielman tulosten tiivistämisen lisäksi viitata aiempaan tutkimukseen. Eri tutkimuksilla saatuja päällekkäisyyksiä ei kuitenkaan voida suoraan vertailla toisiinsa. Vaikka tuloksia ei voida suoraan vertailla toisiinsa, on mahdollista vertailla eri tutkimusten tuloksista nähtäviä suuntauksia toisiinsa. Niinpä seuraavassa Katzerin ym. (1982) tutkimuksen ja tämän tutkielman vertailussa tarkoituksena ei ole niissä esitettyjen päällekkäisyyden prosenttiosuuksien suora vertaaminen toisiinsa. Oleellista on sen sijaan sen tarkasteleminen, toteutuuko tässä tutkielmassa Katzerin ym. (1982) havainto, jonka mukaan tulosjoukkojen päällekkäisyys kasvaa, kun tarkastellaan aina vain relevantimpien dokumenttien välistä päällekkäisyyttä.

Katzer ym. (1982, 269) raportoi, että yhteisten relevanttien dokumenttien osuus kaikista toisiinsa verrattavien tulosjoukkojen relevanteista oli keskimäärin 29 prosenttia täystäsmäyttävää järjestelmää käyttäneessä tutkimuksessa. Yhteisten erittäin relevanttien osuus kaikista erittäin relevanteista dokumenteista oli puolestaan keskimäärin 35 prosenttia. Tutkimuksen tulosten mukaan tulosjoukkojen välinen päällekkäisyys kasvoi, kun tarkasteltavana olevien dokumenttien määrä pieneni tai tarkemmin sanottuna, kun relevanttius arvioitiin tiukemmin. (Katzer ym. 1982, 265–269.) Katzerin ym. (1982) tulokset vastaavat lähinnä tämän tutkielman Yhteisten relevanttien osuus kaikista relevanteista

– kohdassa liberaalilla ja tiukalla relevanssitasolla esitettyjä tuloksia, joten seuraavaksi tarkastellaan niitä.

Ensimmäiseksi tarkastellaan Taulukon 15 englanninkielisen aineiston keskiarvoja. Yhteisten relevanttien osuus kaikista relevanteista on liberaalilla relevanssitasolla keskimäärin 73 prosenttia ja tiukalla relevanssitasolla keskimäärin 64 prosenttia, kun tarkastellaan perusmuotoisen ja karsitun kyselysarjan välisiä päällekkäisyyksiä (ks. Taulukko 15). Tarkasteltaessa Taulukon 15 Yhteisten relevanttien osuus kaikista relevanteista – sarakkeissa esitettyjä keskiarvoja nähdään, että keskiarvo ei vaihtele paljoakaan relevanssitasolta toiselle. Tulosjoukkojen keskimääräinen päällekkäisyys ei näytä siis nousevan eikä laskevan, vaan se pysyy melko samanlaisena eri relevanssitasoilla.

Suomenkielisessä aineistossa keskiarvot ovat seuraavanlaisia. Yhteisten relevanttien osuus kaikista relevanteista on liberaalilla relevanssitasolla keskimäärin 89 prosenttia ja tiukalla relevanssitasolla keskimäärin 83 prosenttia, kun tarkastellaan ositetun perusmuotoisen ja perusmuotoisen kyselysarjan välisiä päällekkäisyyksiä (ks. Taulukko 16). Yhteisten relevanttien osuus kaikista relevanteista on sekä liberaalilla että tiukalla relevanssitasolla keskimäärin 61 prosenttia, kun tarkastellaan perusmuotoisen ja karsitun kyselysarjan välisiä päällekkäisyyksiä (ks. Taulukko 17). Yhteisten relevanttien osuus kaikista relevanteista on liberaalilla relevanssitasolla keskimäärin 59 prosenttia ja tiukalla relevanssitasolla keskimäärin 57 prosenttia, kun tarkastellaan ositetun perusmuotoisen ja karsitun kyselysarjan välisiä päällekkäisyyksiä (ks. Taulukko 18). Yhteenvetona voidaan todeta, että yhteisten relevanttien keskimääräiset osuudet kaikista relevanteista pysyvät melko samanlaisina eri relevanssitasoilla niin englannin- kuin suomenkielisessä aineistossa. Niinpä tämä tutkielma poikkeaa Katzerin ym. (1982) tutkimuksesta, sillä tässä tutkielmassa päällekkäisyys ei juuri muutu, vaikka toisiinsa verrattavien tulosjoukkojen tarkasteltavana olevien dokumenttien määrä tai relevanttiuden aste muuttuisi.

Seuraavassa esitetään syitä sille, miksi äskeinen Katzerin ym. (1982) havainto kasvavasta päällekkäisyydestä ei toteutunut tässä tutkielmassa. Ensinnäkin Katzerin ym. (1982) tutkimus ja tämä tutkielma eroaa toisistaan niissä käytettyjen testikokoelmien, tiedonhakujärjestelmien ja kyselyiden osalta. Nämä samat syyt ovat myös sen taustalla, miksi tällä tutkielmalla ja aiemmilla täystäsmäytystä käyttäneillä tutkimuksilla saatuja tuloksia on ylipäätään vaikea vertailla toisiinsa. Tarkemmin näistä kolmesta ja niiden vaikutuksesta kerrotaan alla.

Eri tutkimuksilla saatujen päällekkäisyyksien vertailun tekee vaikeaksi ensinnäkin se, että eri tutkimuksissa on käytetty erilaisia testikokoelmia. Monesti eri testikokoelmat poikkeavat toisistaan sisältämänsä dokumenttimäärän perusteella. Usein myös eri testikokoelmien saantikannat ovat erikokoisia. Tätä kautta tutkimuksessa käytetty testikokoelma vaikuttaa kyselyillä saatavien tulosjoukkojen kokoon ja tulosjoukkoon sisältyvien relevanttien dokumenttien määrään. Tämä hankaloittaa eri testikokoelmia käyttävien tutkimusten tuloksellisuuksien vertailua ja myös eri testikokoelmia käyttävien tutkimusten tulosjoukkojen päällekkäisyyksien vertailua. Toisekseen eri tutkimuksilla saatujen päällekkäisyyksien vertailua hankaloittaa se, että osa tutkimuksista on tehty täystäsmäyttävää ja osa osittaistäsmäyttävää tiedonhakujärjestelmää käyttäen. Kun tulosjoukkojen pareittaista vertailua tehdään täystäsmäyttävässä järjestelmässä, toisiinsa verrattavat tulosjoukot eivät välttämättä ole samankokoiset. Sen sijaan osittaistäsmäyttävässä järjestelmässä toisiinsa verrattavat tulosjoukot ovat aina keskenään samankokoiset. Tästä johtuen täystäsmäyttävää järjestelmää käytettäessä tulosjoukkojen päällekkäisyys lasketaan eri tavalla kuin osittaistäsmäyttävää järjestelmää käytettäessä. Näin ollen eri täsmäytysmenetelmää käyttäneillä tutkimuksilla havaittuja päällekkäisyyksiä on vaikea vertailla toisiinsa niiden erilaisesta laskutavasta johtuen. Kolmanneksi eri tutkimukset poikkeavat toisistaan niissä käytettyjen kyselyversioiden osalta. Tutkimuksessa käytettävät kyselyversiot laaditaan tutkimuksen hakuaiheiden pohjalta käyttöön valittua kyselyversion laatimistapaa käyttäen. Jos toisiinsa verrattavissa tutkimuksissa on käytetty toisistaan poikkeavia kyselyiden laatimistapoja, on eri tutkimuksilla saatujen tulosten ja päällekkäisyyksien vertailu vaikeaa.

Kerrottakoon Katzerin ym. (1982) tutkimuksen ja tämän tutkielman kyselyversioiden eroista seuraavaa. Tässä tutkimuksessa päällekkäisyyteen tai tuloksellisuuteen eivät vaikuttaneet eri kyselyversioissa käytetyt erilaiset ilmaisut, sillä sekä karsittu että perusmuotoinen kyselyversio sisälsi samat ilmaisut. Ilmaisut esiintyivät kussakin kyselyversiossa siinä muodossa, jonka ne olivat saaneet joko perusmuoto-ohjelman tai karsinta-algoritmin käsittelyn seurauksena. Sen sijaan Katzerin ym. (1982, 263) tutkimuksessa eri kyselyversiot sisälsivät eri ilmaisuja, sillä toisessa versiossa käytettiin hakuavaimina vapaatekstisanoja ja toisessa versiossa asiasanoja kontrolloidusta sanastosta. Asiasanojen määrä on vähäisempi ja muoto kontrolloidumpi kuin vapaatekstisanojen ja muun muassa siksi eri versioiden hakuavain valinnat ja sen myötä tulosjoukot eivät juurikaan olleet päällekkäisiä. Kun eri kyselyversiot sisältävät keskenään eri ilmaisuja, eroavat toisen kyselyversion hakuavaimet verrokkina toimivan toisen kyselyversion hakuavaimista spesifisyydeltään, tyhjentävyydeltään ja homonymial-

taan (Katzer ym. 1982, 273). Katzerin ym. (1982, 273) mukaan on todennäköistä, että tämä spesifisyys, tyhjentävyys ja homonymia vaikuttavat sellaisten tutkimusten tuloksiin, jotka tutkivat eri versioilla saatujen tulosjoukkojen päällekkäisyyksiä.

Tutkielmassa on aiemmin kerrottu, että kaksi erilaista menetelmää voidaan katsoa samankaltaisiksi vasta, kun ne ovat tuloksellisuudeltaan samanlaiset ja ne noutavat samat dokumentit (Das-Gupta & Katzer 1983, 106). Englanninkielisessä aineistossa karsinnan ja perusmuotoistamisen tuloksellisuuserot olivat pienet (1–1,5 prosenttiyksikköä) ja päällekkäisyys melko suurta vertailtaessa kokonaisia tulosjoukkoja (70–74 %). Suomenkielisessä aineistossa karsinnan ja perusmuotoistamisen tuloksellisuuserot olivat suurehkot (3,9–8 prosenttiyksikköä) ja päällekkäisyys keskimääräistä (53–61 %) vertailtaessa kokonaisia tulosjoukkoja. Äskeisen perusteella voidaan sanoa, että englanninkielisessä tekstitiedonhaussa karsinta ja perusmuotoistaminen ovat keskenään samankaltaisemmat menetelmät kuin suomenkielisessä tekstitiedonhaussa.

9 Johtopäätökset

Muistamisen arvoisia asioita tässä tutkielmassa ovat erityisesti seuraavat seikat. Päällekkäisyys näyttää olevan suurinta niiden kyselysarjojen välillä, jotka ovat kyselysarjoina samankaltaisimmat. Päällekkäisyys noudattelee tuloksellisuutta siten, että päällekkäisyys on suurinta niiden menetelmien välillä, jotka ovat tuloksellisuudeltaan samankaltaisimmat ja pienintä niiden menetelmien välillä, joiden tuloksellisuuserot ovat suurimmat.

Saadut tulokset ovat sovellettavissa tekstitiedonhakuun haettaessa kyselyillä, jotka on muodostettu käyttämällä laadinnassa sananmuotojen ohjelmallista käsittelyä siten, että eri kyselyversioiden hakuavaimet poikkeavat toisistaan vain hieman. Tulokset ovat englannin- ja suomenkielisen tekstitiedonhaun näkökulmasta suuntaa antavia. Testatut kielet edustavat ääripäitä, sillä toinen on ominaisuuksiltaan morfologisesti rikas ja toinen morfologisesti yksinkertainen kieli, mutta sen perusteella ei voida päätellä sitä, millainen on tässä tutkitun asian laita muissa morfologisesti rikkaissa tai yksinkertaisissa kielissä.

Päällekkäisyyttä voidaan tarkastella kahdesta näkökulmasta, joista ensimmäistä tarkastellaan tässä kappaleessa ja toista seuraavassa kappaleessa. Ensinnäkin päällekkäisyyttä on tarkasteltu keskittymällä vähäiseen päällekkäisyyteen ja vähäisen päällekkäisyyden merkitykseen. Tämä näkökulma on hyvin usein liitetty yhdistelyn tutkimiseen. Jotta tätä näkökulmaa edustava jatkotutkimus olisi menestyksekäs, tulee seuraavien reunaehtojen täyttyä. Tulosjoukkojen välisen päällekkäisyyden tulee olla vähäistä, sillä yhdistely ei paranna tiedonhaun tulosta, jos toisiinsa verrattavilla tulosjoukoilla on paljon yhteisiä dokumentteja. Jotta yhdistely toimisi, tulee myös tulosjoukkojen tuloksellisuuden olla melko hyvä.

Toisekseen päällekkäisyyttä voidaan tarkastella evidenssinä relevanssista. Tässä toisessa näkökulmassa tulosjoukkojen päällekkäisyyttä ajatellaan hyödynnettävän siten, että dokumenttien yhteisesiintymistä käytetään evidenssinä relevanssista. Toisin sanoen tulosjoukkojen päällekkäisyyden avulla vahvistetaan relevanttien dokumenttien löytymisen todennäköisyyttä. Tällä näkökulmalla ei siis ole mitään tekemistä yhdistelyn eikä vähäisen päällekkäisyyden kanssa. Asia on pikemminkin päinvastoin, sillä mitä enemmän toisiinsa verrattavilla kokonaisilla tulosjoukoilla on keskenään yhteisiä dokumentteja, sitä suuremmalla todennäköisyydellä kyseiset dokumentit ovat relevantteja. Jotta päällekkäisyys toimisi relevanssin evidenssinä, täytyy yhteisten dokumenttien olla relevantteja. Tässä tutkielmassa saatujen tulosten perusteella voisi ajatella, että liberaalilla relevanssitasolla ja alhaisilla vertailupisteillä tämä toimii kohtuullisen luotettavasti, mutta tiukalla relevanssitasolla luotettavuus huononee.

10 Lähdeluettelo

Airio, E., Keskustalo, H., Hedlund, T. & Pirkola, A. 2003. Multilingual experiments of UTA at CLEF 2003: the impact of different merging strategies and word normalizing tools. Teoksessa C. Peters (toim.) Working Notes for the CLEF 2003 Workshop, 21–22 August, Trondheim, Norway. Saatavilla pdf-muodossa: <URL: http://www.clef-campaign.org/2003/WN_web/02.pdf>. (Viitattu 10.9.2005).

Airio, E., Keskustalo, H., Hedlund, T. & Pirkola, A. 2004. The impact of word normalization methods and merging strategies on multilingual IR. Teoksessa Comparative Evaluation of Multilingual Information Access Systems. Lecture Notes in Computer Science 3237/2004. Heidelberg: Springer-Verlag, 74–84. Saatavilla SpringerLink tietokannasta lisenssiä vastaan: <URL: http://springerlink.metapress.com>. (Viitattu 12.5.2006).

Alaterä, A. & Halttunen, K. 2003. Tiedonhaun perusteet - osa lukutaitoa? Saatavilla www-muodossa: <URL: http://www.internetix.ofw.fi/opinnot/opintojaksot/0viestinta/informaatiotutkimus/po2/>. (Viitattu 12.2.2006).

Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swan, R. & Xu, J. 1997. INQUERY does battle with TREC-6. Teoksessa E. M. Voorhees & D. K. Harman (toim.) NIST Special Publication 500–240: the Sixth Text REtieval Conference (TREC-6), 169–206. Saatavilla PostScript-muodossa: <URL: http://trec.nist.gov/pubs/trec6/papers/umass-trec6.ps.gz>. (Viitattu 10.1.2006).

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

Belkin, N. J., Cool, C., Croft, W. B. & Callan, J. P. 1993. The effect of multiple query representations on information retrieval system performance. Teoksessa R. Korfhage, E. Rasmussen & P. Willett (toim.) Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, Pittsburg, PA, June 27–July 1, 1993. New York: ACM Press, 339–

346. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 5.12.2005).

Braschler, M. 2004. Combination approaches for multilingual text retrieval. Information Retrieval 7 (1–2), 183–204.

Braschler, M. & Ripplinger, B. 2003. Stemming and decompounding for German text retrieval. Teoksessa Advances in Information Retrieval. Lecture Notes in Computer Science 2633/2003. Heidelberg: Springer, 177–192. Artikkeliin viitattu artikkelissa Kettunen, K., Kunttu, T. & Järvelin, K. 2005. To stem or lemmatize a highly inflectional language in probabilistic IR environment? Journal of Documentation 61 (4), 476–496.

Callan, J. P., Croft, W. B. & Harding, S. M. 1992. The INQUERY retrieval system. Saatavilla PostScript-muodossa: <URL: http://Santana.uni-muenster.de/Library/Virtual/InformationRetrieval/rqs_18.11.ps>. (Viitattu 17.5.2005). Julkaistu myös painettuna teoksessa Proceedings of the Third International Conference on Database and Expert Systems Applications, Valencia, Spain. Wien: Springer-Verlag, 78–83.

Creutz, M. 2006. Induction of the morphology of natural language: unsupervised morpheme segmentation with application to automatic speech recognition. Helsinki University of Technology. Dissertations in Computer and Information Science. Väitöskirja. Saatavilla pdf-muodossa: <URL: http://lib.tkk.fi/Diss/2006/isbn9512282119/isbn9512282119.pdf>. (Viitattu 11.10.2006).

Das-Gupta, P. & Katzer, J. 1983. A study of the overlap among document representations. Teoksessa Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval, Bethesda, MD, June 6–8, 1983. New York: ACM Press, 106–114. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 11.10.2006).

Frakes, W. B. & Fox, C. J. 2003. Strength and similarity of affix removal stemming algorithms. ACM SIGIR Forum 37 (1), 26–30. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 5.12.2005).

Hakulinen, A. & Ojanen, J. 1993. Kielitieteen ja fonetiikan termistöä. 3. painos. Suomalaisen Kirjallisuuden Seuran toimituksia 324. Helsinki: Suomalaisen Kirjallisuuden Seura.

Harman, D. 1991. How effective is suffixing? Journal of the American Society for Information Science 42 (1), 7–15.

Harman, D. 1987. A failure analysis on the limitations of suffixing in an online environment. Teoksessa C. T. Yu & C. J. van Rijsbergen (toim.) Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information Retrieval, New Orleans, LA, June 3–5, 1987. New York: ACM Press, 102–108. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 10.12.2005).

Hollink, V., Kamps, J., Monz, C. & de Rijke, M. 2004. Monolingual document retrieval for European languages. Information Retrieval 7 (1-2), 33–52.

Hood, W. W. & Wilson, C. S. 2003. Overlap in bibliographic databases. Journal of the American Society for Information Science and Technology 54 (12), 1091–1103.

Hull, D. A. 1996. Stemming algorithms: a case study for detailed evaluation. Journal of the American Society for Information Science 47 (1), 70–84.

Häkkinen, K. 1994. Kielitieteen perusteet. Tietolipas 133. Helsinki: Suomalaisen Kirjallisuuden Seura.

Iivonen, M. 1989. Indeksointituloksen riippuvuus indeksointiympäristöstä. Tampereen yliopiston kirjastotieteen ja informatiikan laitoksen tutkimuksia 26.

Ingwersen, P. 1992. Information retrieval interaction. London: Taylor Graham. Saatavilla www-muodossa: <URL: http://www.db.dk/pi/iri//>. (Viitattu 14.4.2006).

Ingwersen, P. 1994. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. Teoksessa W. B. Croft & C. J. van Rijsbergen

(toim.) Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland, July 3–6, 1994. New York: Springer-Verlag, 101–110. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 8.11.2006).

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

Järvelin, K. & Kekäläinen, J. 2002. Tiedonhaun menetelmät. Saatavilla www-muodossa: <URL: www.internetix.fi/opinnot/opintojaksot/0viestinta/informaatiotutkimus/po4/>. (Viitattu 26.4.2004).

Karlsson, F. 1998. Yleinen kielitiede. Uud. laitos. Helsinki: Yliopistopaino.

Karlsson, F. 1983. Suomen kielen äänne- ja muotorakenne. Helsinki: WSOY.

Katzer, J., McGill, M. J., Tessier, J. A., Frakes, W. & DasGupta, P. 1982. A study of the overlap among document representations. Information Technology : Research and Development 2, 261–274.

Kettunen, K. 2005. Sijamuodot haussa: tarvitseeko kaikkea hakutermien morfologista vaihtelua kattaa? Tampereen yliopisto. Informaatiotutkimuksen laitos. Sivuainetutkielma.

Kettunen, K., Kunttu, T. & Järvelin, K. 2005. To stem or lemmatize a highly inflectional language in probabilistic IR environment? Journal of Documentation 61 (4), 476–496.

Klammer, T. P. & Schulz, M. R. 1992. Analyzing English grammar. Boston: Allyn and Bacon.

Korhonen, R., Vilkuna, M. & Vihtari, J. 2005. Sananselityksiä: ison suomen kieliopin termejä. Kotimaisten kielten tutkimuskeskuksen verkkojulkaisuja 1. Saatavilla www-muodossa: <URL: http://www.kotus.fi/verkkojulkaisut/julk1/>. (Viitattu 9.7.2006).

Koskenniemi, K. 1985. An application of the two-level model to Finnish. Teoksessa F. Karlsson (toim.) Computational morphosyntax: report on research 1981–84. University of Helsinki. Publications of the Department of general linguistics 13, 19–42.

Krott, A., Baayen, R. H. & Schreuder, R. 2001. Analogy in morphology: modeling the choice of linking morpheme in Dutch. Linguistics 39 (1), 51–93. Artikkeliin viitattu artikkelissa Hollink, V., Kamps, J., Monz, C. & de Rijke, M. 2004. Monolingual document retrieval for European languages. Information Retrieval 7 (1–2), 33–52.

Krovetz, R. 2000. Viewing morphology as an inference process. Artificial Intelligence 118 (1–2), 277–294. Artikkeliin viitattu artikkelissa Kettunen, K., Kunttu, T. & Järvelin, K. 2005. To stem or lemmatize a highly inflectional language in probabilistic IR environment? Journal of Documentation 61 (4), 476–496.

Krovetz, R. J. 1995. Word sense disambiguation for large text databases. University of Massachusetts Amherst. Department of Computer Science. Vaitöskirja. Väitöskirjaan viitattu artikkelissa Porter, M. F. 2001. Snowball: a language for stemming algorithms. Saatavilla www-muodossa: <URL: http://snowball.tartarus.org/texts/introduction.html>. (Viitattu 17.11.2005).

Kunttu, T. 2003. Perus- ja taivutusmuotohakemiston tuloksellisuus todennäköisyyksiin perustuvassa tiedonhakujärjestelmässä. Tampereen yliopisto. Informaatiotutkimuksen laitos. Pro gradu -tutkielma.

Laaksonen, K. & Lieko, A. 2003. Suomen kielen äänne- ja muoto-oppi. 4. uud. painos. Helsinki: Finn Lectura.

Laalo, K. 1990. Säkeistä patoihin: suomen kielen monitulkintaiset sananmuodot. Suomi 154. Helsinki: Suomalaisen Kirjallisuuden Seura.

Laine, H. 2005. Tietokantojen perusteet. Saatavilla pdf-muodossa: <URL: http://www.cs.helsinki.fi/u/laine/tikape/s05/pdf/johdanto_c.pdf>. (Viitattu 1.12.2005).

Lee, J. H. 1996. Combining multiple evidence from different relevance feedback methods. University of Massachusetts Amherst, Center for Intelligent Information Retrieval. CIIR Technical Report. Saatavilla PostScript-muodossa: <URL: http://ciir.cs.umass.edu/info/psfiles/irpubs/ir87.ps.gz>. (Viitattu 9.11.2005).

Lepäsmaa, A-L., Lieko, A. & Silfverberg, L. 1996. Miten sanoja johdetaan: suomen kielen johto-oppia. Helsinki: Finn Lectura.

Lexicon of linguistics 2001. Kerstens J., Ruys E. & Zwarts J. (toim.) Utrecht University. Saatavilla www-muodossa: <URL: http://www2.let.uu.nl/UiL-OTS/Lexicon/>. (Viitattu 10.12.2006).

Matthews, P. H. 1991. Morphology. 2. painos. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Mattila, I. & Mattila, M. 1997. Lukion englannin kielioppi. Helsinki: Otava.

McAlester, G., Nyberg, S., Oksanen, L., Olkkonen, R., Pylkki, L., Pääkkönen, P. & Tuokko, E. 1992. As a rule: englannin keskeinen kielioppi. Helsinki: WSOY.

McCain, K. W. 1989. Descriptor and citation retrieval in the medical behavioral sciences literature: retrieval overlaps and novelty distribution. Journal of the American Society for Information Science 40 (2), 110–114.

McGill, M., Koll, M. & Noreault, T. 1979. An evaluation of factors affecting document ranking by information retrieval systems. Syracuse University. School of Information Studies.

Paice, C. D. 1994. An evaluation method for stemming algorithms. Teoksessa W. B. Croft & C. J. van Rijsbergen (toim.) Proceeding of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland, July 3–6, 1994. New York: Springer-Verlag, 42–50. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 3.10.2005).

Pao, M. L. 1984. Semantic and pragmatic retrieval. Teoksessa Proceedings of the 47th ASIS Annual Meeting, Philadelphia, Pennsylvania, October 21–25, 1984. New York: Knowledge Industry Publications, 134–136. Artikkeliin viitattu artikkelissa McCain, K. W. 1989. Descriptor and citation retrieval in the medical behavioral sciences literature: retrieval overlaps and novelty distribution. Journal of the American Society for Information Science 40 (2), 110–114.

Pao, M. L. 1994. Relevance odds of retrieval overlaps from seven search fields. Information Processing & Management 30 (3), 305–314.

Penttilä, A. 1975. Homonüümiast, eriti soome keelt silmas pidades. Congressus Tertius Internationalis Fenno-Ugristarum. Tallinnae Habitus 17.–23. VIII 1970. Pars I, 322–326. Kongressiesitelmään viitattu teoksessa Laalo, K. 1990. Säkeistä patoihin: suomen kielen monitulkintaiset sananmuodot. Suomi 154. Helsinki: Suomalaisen Kirjallisuuden Seura.

Penttilä, A. 2002. Suomen kielioppi. 3. muuttam. painos. Vantaa: Dark.

Plag, I. 2003. Word-formation in English. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Porter, M. F. 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Porter, M. F. 2001. Snowball: a language for stemming algorithms. Saatavilla www-muodossa: <URL: http://snowball.tartarus.org/texts/introduction.html>. (Viitattu 17.11.2005).

Popovič, M. & Willett, P. 1992. The effectiveness of stemming for natural-language access to Slovene textual data. Journal of the American Society for Information Science 43 (5), 384–390.

Rajashekar, T. B. & Croft, W. B. 1995. Combining automatic and manual index representations in probabilistic retrieval. Journal of the American Society for Information Science 46 (4), 272–283.

Salton, G. & McGill, M. 1983. Introduction to modern information retrieval. New York: McGraw-Hill.

Saracevic, T. 1996. Relevance reconsidered ´96. Teoksessa P. Ingwersen & N. O. Pors (toim.) Proceedings CoLIS 2: Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, October 13–16, 1996. København: The Royal School of Librarianship, 201–218.

Saracevic, T. & Kantor, P. 1988. A study of information seeking and retrieving. III. Searchers, searches, and overlap. Journal of the American Society for Information Science 39 (3), 197–216. Saatavilla pdf-muodossa: <URL: http://www.scils.rutgers.edu/~tefko/JASIS1988part3.pdf>. (Viitattu 2.12.2005).

Saukkonen, P. 1973. Suomen kielen yhdyssanojen rakenne. Teoksessa Commentationes Fenno-Ugricae in honorem Erkki Itkonen sexagenarii die XXVI mensis aprilis anno MCMLXXIII. Erkki Itkonen 60 v. Suomalais-Ugrilaisen Seuran Toimituksia 150. Helsinki: Suomalais-Ugrilainen Seura, 332–339.

Savolainen, E., Haakana, M., Lieko, A., Muikku-Werner, P. & Mäntynen, A. 1997. Hyperkielioppi: multimediasovellus suomen kieliopista ja murteista sekä kirjakielen kehityksestä ja huollosta [CD-ROM]. Helsinki: Finn Lectura.

Sormunen, E. 2000. A method for measuring wide range performance of Boolean queries in full-text databases. Tampereen yliopisto. Acta Universitatis Tamperensis 748. Väitöskirja.

Sormunen, E. 2002. Liberal relevance criteria of TREC – Counting on negligible documents? Teoksessa Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information Retrieval, Tampere, Finland, August 11–15, 2002. New York: ACM Press, 324–330. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 2.12.2005).

Sormunen, E., Kekäläinen, J., Koivisto, J. & Järvelin, K. 2001. Document text characteristics affect the ranking of the most relevant documents by expanded structured queries. Journal of Documentation 57 (3), 358–376.

Sparck Jones, K. 1974. Automatic indexing. Journal of Documentation 30 (4), 393–432.

Sparck Jones, K. 1990. Retrieving information or answering questions? The British Library Annual Research Lecture 8. London: British Library. Artikkeliin viitattu artikkelissa Ingwersen, P. 1994. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. Teoksessa W. B. Croft & C. J. van Rijsbergen (toim.) Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland, July 3–6, 1994. New York: Springer-Verlag, 101–110. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 8.11.2006).

Text Retrieval Conference (TREC). 2004. Overview. Saatavilla www-muodossa: <URL: http://trec.nist.gov/overview.html>. (Viitattu 21.1.2005).

Text Retrieval Conference (TREC). 2006. Data – English relevance judgements. Saatavilla www-muodossa: <URL:http://trec.nist.gov/data/reljudge_eng.html>. (Viitattu 25.4.2007).

Tomlinson, S. 2003. Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003. Teoksessa C. Peters (toim.) Working Notes for the CLEF 2003 Workshop, 21–22 August, Tronheim, Norway. Saatavilla pdf-muodossa: <URL: http://www.clef-campaign.org/2003/WN_web/19.pdf>. (Viitattu 14.3.2005).

Turtle, H. & Croft, W. B. 1990. Inference networks for document retrieval. Teoksessa J-L. Vidick (toim.) Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval, Brussels, Belgium, September 5–7, 1990. New York: ACM Press, 1–24. Saatavilla ACM-portaalista lisenssiä vastaan: <URL: http://portal.acm.org>. (Viitattu 5.12.2005).

Vannest, J., Bertram, R., Järvikivi, J. & Niemi, J. 2002. Counterintuitive cross-linguistic differences: more morphological computation in English than in Finnish. Journal of Psycholinguistic Research 31 (2), 83–105.

Voorhees, E. M. 2003. Overview of TREC 2003. Teoksessa E. M. Voorhees & Lori P. Buckland (toim.) NIST Special Publication 500–255: The Twelfth Text REtrieval Conference (TREC 2003) 1–13. Saatavilla pdf-muodossa: <URL: http://trec.nist.gov/pubs/trec12/papers/OVERVIEW.12.pdf>. (Viitattu 30.3.2005).

Liitteet

Liite 1: TRECin perusmuotoisten ja karsittujen kyselysarjojen saanti-tarkkuusarvot

TAULUKKO 19. (TREC) Perusmuotoisen ja karsitun kyselysarjan saanti-tarkkuusarvot liberaalilla, normaalilla ja tiukalla relevanssitasolla.

	Liberaali relevanssitaso			**Normaali relevanssitaso**			**Tiukka relevanssitaso**	
Saan-titaso	Perusmuotoi-nen kysely-sarja	Karsittu kyse-lysarja		Perusmuotoi-nen kysely-sarja	Karsittu kyse-lysarja		Perusmuotoi-nen kysely-sarja	Karsittu kyse-lysarja
10	44,8	48,8		50,9	52,5		40,5	42,8
20	34,4	35,4		37,8	38,8		31,4	34,9
30	26,2	28		27,4	28		27,9	30
40	20,2	21,2		23,5	24,8		20,2	21,1
50	15,2	16,1		19,5	21,2		19,3	19,8
60	10,2	12,1		15,3	16,5		14,6	14,7
70	7,1	8,2		12,3	13,4		10,9	12,2
80	3,6	5,9		8,2	10,2		10,4	11,4
90	1,9	2,5		5,2	5,4		9,1	9,5
100	1,3	1,6		3,3	3,5		9	9,3

Liite 2: Hakuaiheet

Englanninkieliset hakuaiheet:

q351 What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?

q353 Identify systematic explorations and scientific investigations of Antarctica, current or planned.

q355 Identify documents discussing the development and application of spaceborne ocean remote sensing.

q358 What role does blood-alcohol level play in automobile accident fatalities?

q360 What are the benefits, if any, of drug legalization?

q362 Identify incidents of human smuggling.

q364 Identify documents discussing cases where rabies have been confirmed and what, if anything, is being done about it.

q365 What effects have been attributed to El Nino?

q372 Identify documents that discuss the growth of Native American casino gambling.

q373 Identify documents that discuss the concerns of the United States regarding the export of encryption equipment.

q377 Identify documents that discuss the renewed popularity of cigar smoking.

q378 Identify documents that discuss opposition to the introduction of the euro, the European currency.

q384 Identify documents that discuss the building of a space station with the intent of colonizing the moon.

q385 Identify documents that discuss the current status of hybrid automobile engines, (i.e., cars fueled by something other than gasoline only).

q387 Identify documents that discuss effective and safe ways to permanently handle long-lived radioactive wastes.

q388 Identify documents that discuss the use of organic fertilizers (composted sludge, ash, vegetable waste, microorganisms, etc.) as soil enhancers.

q392 What are the applications of robotics in the world today?

q393 Identify documents that discuss mercy killings.

q396 Identify documents that discuss sick building syndrome or building-related illnesses.

q399 Identify documents that discuss the activities or equipment of oceanographic vessels.

q400 What measures are being taken by local South American authorities to preserve the Amazon tropical rain forest?

Suomenkieliset hakuaiheet:

q1. George Bushin ja Mihail Gorbatshovin tapaaminen Helsingissä syyskuussa 1990. Neuvotteluissa käsitellyt asiat sekä tehdyt päätökset ja sopimukset.

q2. Etelä-Amerikan velkakriisi. Miten velkaantumisongelma on kehittynyt? Miten ongelmaa on pyritty ratkaisemaan?

q3. Metsäteollisuuden polkumyyntisyytteet USA:ssa. Kiinnostavaa suomalaisten paperinviejien kohtalo. Polkumyyntisyytösten sisältö, oikeudenkäynnin tulokset.

q4. Jyväskylän kaupungin ja maalaiskunnan kuntaliitoshanke. Halutaan kartoittaa liitoshankkeen kannattajien ja vastustajien mielipiteitä ja perusteluja. Arviot liitoksen taloudellisista vaikutuksista (mm. porkkanaraha).

q6. Varsovan liiton lakkauttaminen. Mitä tahansa muutosprosessista, eri maiden suhtautumisesta, päätöksistä, jne.

q7. Neuvostoliiton Liettuaan kohdistama taloussaarto keväällä 1990. Mitä toimia taloussaartoon liittyi ja miten se näkyi Liettuassa? Saarron lopettamiseen johtaneet tapahtumat.

q8. Irakin joukkotuhoaseiden hävittäminen. Irakin on Persianlahden sodan aseleposopimuksen mukaan luovuttava kemiallisista, biologisista ja ydinaseista ja niiden tuotantotekniikasta. YK vastaa aseiden inventoinnista ja hävittämisestä. Miten tehtävän suoritus on onnistunut?

q9. OPEC:n öljyn hintaa ja tuotantomääriä koskevat päätökset.

q10. Presidentti Illiescun hallituksen avuksi kutsumien kaivosmiesten väkivaltaisuudet oppositiota vastaan Bukarestissa. Taustatietoja tapahtumista, uhreista ja jälkiselvittelystä.

q11. Namibian itsenäistymiseen liittynyt YK:n rauhanturvaoperaatio. Tietoja operaation valmistelusta, siihen liittyneistä tapahtumista sekä UNTAG-joukkojen ja sen suomalaispataljoonan toiminnasta.

q12. EY:n parlamentin asema yhteisön päätöksenteossa. Halutaan selvittää EY:n parlamentin asema suhteessa komissioon ja ym. toimielimiin. Mitä muutoksia nykyiseen on haluttu ja ketkä ovat halunneet? Miten demokraattinen kontrolli toimii EY:ssä?

q13. Carl Bildt ja pohjoismainen yhteistyö. Bildtin pohjoismaista yhteistyötä koskevat lausunnot. Mitä erityistä Bildt on sanonut Ruotsin ja Suomen yhteistyöstä?

q14. Jugoslavian presidenttineuvoston toimintaa koskevat uutiset. Erityisesti tiedot istunnoista ja niissä tehdyistä päätöksistä.

q15. Länsi- ja Itä-Saksan sekä miehittäjävaltojen (Yhdysvallat, Iso-Britannia, Ranska ja Neuvostoliitto) välillä käytiin 2+4-neuvotteluja Saksojen yhdistymisestä. Mitkä olivat keskeisimmät ratkaistavat

kysymykset? Mitä erityisiä riitakysymyksiä nousi esiin? Mitä olennaista syntyneisiin sopimuksiin sisältyy?

q17. Valmetin traktori- ja kuljetusvälinetuotannon kannattavuus. Kuljetusvälinetoimialaan lasketaan kuuluviksi metsä- ja siirtokoneet sekä kiskokaluste (mm. Transtech). Osakkuudet henkilö- ja kuorma-autoteollisuudessa jätetään tarkastelun ulkopuolelle.

q20. Tampellan irtisanomiset. Tavoitteena koota tietoja Tampella-konserniin kuuluvien yhtiöiden suorittamista irtisanomisista. Tietoja lomautuksista ja lyhennetyistä työviikoista ei tarvita.

q21. Keran ja KTM:n investoinnit matkailuun. Tietoja matkailualan yrityksille myönnetyistä avustuksista ja lainoista (= tässä investointi). Erityisen arvokkaita yhteenvedot.

q22. Neste Oy:n maakaasutoiminta. Halutaan yleiskuva Nesteen maakaasutoiminnoista. Mitä Neste on puuhaillut maakaasun hankinnan (kentät ja tuontisopimukset), jakelun (verkoston rakentaminen) ja markkinoinnin alueilla.

q23. Ydinvoimalaitosten tuottamien radioaktiivisten jätteiden käsittely ja varastointi. Esimerkkejä ongelmista, riskeistä ja sattuneista ydinjätevahingoista.

q24. AIDSin levinneisyys EY-maissa. Miten vakava AIDS-tilanne on näissä maissa? Tietoja esiintymämääristä ja kampanjoista ym. taudin leviämistä ehkäisevistä toimista.

q25. Elintarvikkeiden tuontirajoitusten poiston vaikutus Suomen elintarviketeollisuuteen.

q26. Asuntotuotannon suhdanteet ja suhdannevaihtelut Suomessa (valtakunnallinen taso); erityisesti tilasto- ja ennustetietoja, arvioita (rakentamisesta, ei asuntokaupasta).

q27. Tieliikenteen päästöt Suomessa ja ulkomailla. Miten päästöt ovat kehittyneet ja niiden odotetaan kehittyvän (mm. lainsäädännön vaikutus). Miten merkittävästi katalysaattorien yleistyminen vaikuttaa päästötasoihin? Katalysaattoritekniikka ei sinänsä kiinnosta.

q28. Japanin autoteollisuuden investoinnit Eurooppaan ja tuotannollinen yhteistyö eurooppalaisten autonvalmistajien kanssa. Mihin maihin japanilaisia autotehtaita on suunniteltu, perustettu ja laajennettu? Tuotantomäärät ja -trendit.

q29. Metsäteollisuuden ympäristöinvestoinnit. Rajoitutaan vesiensuojeluun liittyviin investointeihin kemiallisessa metsäteollisuudessa. Sekä varsinaiset puhdistamoinvestoinnit että ympäristöystävällisempien prosessien käyttöönotto.

q30. Kaupan aukioloajat. Halutaan selvittää vähittäiskauppojen aukioloaikojen vapauttamista koskevaa keskustelua. Erityisesti kartoitetaan kaupan järjestöjen ja ammattijärjestöjen kannanottoja ja toimia.

q31. Pakkaukset ympäristönsuojelukysymyksenä. Erityisesti kiinnostavat kulutustavarapakkausten kierrätysjärjestelmät, niiden kehittämiskokeilut, kierrätykseen liittyvä lainsäädäntö eri maissa.

q33. Esko Aho ja Suomen EY-jäsenhakemus. Ahon Suomen EY-jäsenyyden hakemiseen liittyvät mielipiteet, kannanotot ja toimet. Muiden arviot Eskon puheista ja toimista.

q34. Kauko Juhantalon ydinvoimapuheet ja -teot. Juhantalon perustelut 5. ydinvoimalan puolesta. Miten Juhantalo vei ydinvoimalaratkaisua eteenpäin?

q35. Vihreiden tekemät aloitteet, välikysymykset, ehdotukset, puheenvuorot ja äänestyskäyttäytyminen Suomen eduskunnassa. Tarkastelussa sekä ryhmä että yksittäiset kansanedustajat.

Liite 3: Kyselyt

Englanninkieliset perusmuotoiset kyselyt perusmuotohakemistoon

#q351=#sum(information available petroleum exploration south @atlantic near @falkland island);

#q353=#sum(identify systematic exploration scientific investigation @antarctica current plan);

#q355=#sum(identify document discuss development application @spaceborne ocean remote sense);

#q358=#sum(role #0(blood alcohol) level play automobile accident fatality);

#q360=#sum(benefit drug legalization);

#q362=#sum(identify incident human smuggle);

#q364=#sum(identify document discuss case rabies confirm);

#q365=#sum(effect attribute el @nino);

#q372=#sum(identify document discuss growth native @american casino gamble);

#q373=#sum(identify document discuss concern unite state regard export encryption equipment);

#q377=#sum(identify document discuss renew popularity cigar smoke);

#q378=#sum(identify document discuss opposition introduction @euro @european currency);

#q384=#sum(identify document discuss build space station intent colonize moon);

#q385=#sum(identify document discuss current status hybrid automobile engine car fuel gasoline);

#q387=#sum(identify document discuss effective safe way permanent handle #0(long live) radioactive waste);

#q388=#sum(identify document discuss organic fertilizer compost sludge ash vegetable waste micro-organism soil enhancer);

#q392=#sum(application robotics world today);

#q393=#sum(identify document discuss mercy killing);

#q396=#sum(identify document discuss sick build syndrome #0(build relate) illness);

#q399=#sum(identify document discuss activity equipment oceanographic vessel);

#q400=#sum(measure take local south @american authority preserve amazon tropical rain forest);

#q402=#sum(happen field behavioral genetics study relative influence genetic environmental factor individual behavior personality);

#q403=#sum(find information effect dietary intake potassium magnesium fruit vegetable determinant bone mineral density elderly man woman prevent osteoporosis bone decay);

#q405=#sum(unexpected unexplained cosmic event celestial phenomenon radiation supernova out-burst new comet detect);

#q407=#sum(impact poach world various wildlife preserve);

#q408=#sum(tropical storm hurricane typhoon cause significant property damage loss life);

#q410=#sum(involve @schengen agreement eliminate border control western @europe hope accom-plish);

#q414=#sum(sugar @cuba export country import);

#q415=#sum(know drug traffic golden triangle area @burma @thailand @laos meet);

#q416=#sum(status three gorge project);

#q418=#sum(way quilt generate income);

#q420=#sum(widespread carbon monoxide poison global scale);

#q421=#sum(disposal industrial waste accomplish industrial management world);

#q427=#sum(find document discuss damage ultraviolet @uv light sun eye);

#q428=#sum(country china decline birth rate);

#q431=#sum(late development robotic technology);

#q437=#sum(experience residential utility customer follow deregulation gas electric);

#q440=#sum(step take government corporation eliminate abuse child labor);

#q442=#sum(find account selfless heroic act individual small group benefit cause);

#q445=#sum(country unite state consider approve woman clergy person);

#q448=#sum(identify instance weather main contribute factor loss ship sea);

Englanninkieliset karsitut kyselyt karsittuun hakemistoon

#q351=#sum(inform avail petroleum explor south atlant near falkland island);

#q353=#sum(identifi systemat explor scientif investig antarctica current plan);

#q355=#sum(identifi document discuss develop applic spaceborn ocean remot sens);

#q358=#sum(role blood-alcohol level plai automobil accid fatal);

#q360=#sum(benefit drug legal);

#q362=#sum(identifi incid human smuggl);

#q364=#sum(identifi document discuss case rabi confirm done);

#q365=#sum(effect attribut el nino);

#q372=#sum(identifi document discuss growth nativ american casino gambl);

#q373=#sum(identifi document discuss concern unit state regard export encrypt equip);

#q377=#sum(identifi document discuss renew popular cigar smoke);

#q378=#sum(identifi document discuss opposit introduct euro european currenc);

#q384=#sum(identifi document discuss build space station intent colon moon);

#q385=#sum(identifi document discuss current statu hybrid automobil engin car fuel gasolin);

#q387=#sum(identifi document discuss effect safe wai perman handl long-live radioact wast);

#q388=#sum(identifi document discuss organ fertil compost sludg ash veget wast microorgan soil enhanc);

#q392=#sum(applic robot world todai);

#q393=#sum(identifi document discuss merci kill);

#q396=#sum(identifi document discuss sick build syndrom build-relat ill);

#q399=#sum(identifi document discuss activ equip oceanograph vessel);

#q400=#sum(measur taken local south american author preserv amazon tropic rain forest);

#q402=#sum(happen field behavior genet studi rel influenc genet environment factor individu behavior person);

#q403=#sum(find inform effect dietari intak potassium magnesium fruit veget determin bone miner densiti elderli men women prevent osteoporosi bone decai);

#q405=#sum(unexpect unexplain cosmic event celesti phenomena radiat supernova outburst new comet detect);

#q407=#sum(impact poach world variou wildlif preserv);

#q408=#sum(tropic storm hurrican typhoon caus signific properti damag loss life);

#q410=#sum(involv schengen agreement elimin border control western europ hope accomplish);

#q414=#sum(sugar cuba export countri import);

#q415=#sum(known drug traffick golden triangl area burma thailand lao meet);

#q416=#sum(statu three gorg project);

#q418=#sum(wai quilt gener incom);

#q420=#sum(widespread carbon monoxid poison global scale);

#q421=#sum(dispos industri wast accomplish industri manag world);

#q427=#sum(find document discuss damag ultraviolet uv light sun ey);

#q428=#sum(countri china declin birth rate);

#q431=#sum(latest develop robot technolog);

#q437=#sum(experi residenti util custom follow deregul ga electr);

#q440=#sum(step taken govern corpor elimin abus child labor);

#q442=#sum(find account selfless heroic act individu small group benefit caus);

#q445=#sum(countri unit state consid approv women clergi person);

#q448=#sum(identifi instanc weather main contribut factor loss ship sea);

Suomenkieliset perusmuotoiset kyselyt perusmuotohakemistoon

#q1=#sum(george bush mihail gorbatshov tapaaminen helsinki syyskuu 1990 neuvottelu käsitelty asia tehty päätös sopimus);

#q2=#sum(#0(etelä amerikka) velkakriisi velkaantumisongelma kehittyä ongelma pyrkiä ratkaista);

#q3=#sum(metsäteollisuus polkumyyntisyyte usa kiinnostava suomalainen paperinviejä kohtalo polkumyyntisyytös sisältö oikeudenkäynti tulos);

#q4=#sum(jyväskylä kaupunki maalaiskunta kuntaliitoshanke kartoittaa liitoshanke kannattaja vastustaja mielipide perustelu arvio liitos taloudellinen vaikutus porkkanaraha);

#q6=#sum(varsova liitto lakkauttaminen muutosprosessi eri maa suhtautuminen päätös);

#q7=#sum(neuvostoliitto liettua kohdistaa taloussaarto kevät 1990 toimi taloussaarto liittyä näkyä liettua saarto lopettaminen johtaa tapahtuma);

#q8=#sum(irak joukkotuhoase hävittäminen irak persianlahti sota aseleposopimus luopua kemiallinen biologinen ydinase tuotantotekniikka yk vastata ase inventointi hävittäminen tehtävä suoritus onnistua);

#q9=#sum(opec öljy hinta tuotantomäärä koskea päätös);

#q10=#sum(presidentti @illiescun hallitus apu kaivosmies väkivaltaisuus oppositio bukarest taustatieto tapahtuma uhri jälkiselvittely);

#q11=#sum(namibia itsenäistyminen liittyä yk rauhanturvaoperaatio tieto operaatio valmistelu liittyä tapahtuma #0(@untag joukko) suomalaispataljoona toiminta);

#q12=#sum(ey parlamentti asema yhteisö päätöksenteko selvittää ey parlamentti asema suhde komissio toimielin muutos nykyinen haluttu halunnut demokraattinen kontrolli toimia ey);

#q13=#sum(carl @bildt pohjoismainen yhteistyö @bildtin pohjoismainen yhteistyö koskea lausunto eritty @bildt sanoa ruotsi suomi yhteistyö);

#q14=#sum(jugoslavia presidenttineuvosto toiminta koskea uutinen tieto istunto tehty päätös);

#q15=#sum(länsi #0(itä saksa) miehittäjävalta yhdysvalta #0(iso britannia) ranska neuvostoliitto #1(2+4 neuvottelu) saksa yhdistyminen ratkaista kysymys riitakysymys olennainen syntyä sopimus);

#q17=#sum(valmet traktori kuljetusvälinetuotanto kannattavuus kuljetusvälinetoimiala kuuluva metsä siirtokone kiskokaluste @transtech osakkuus henkilö #0(kuorma autoteollisuus) jättää tarkastelu);

#q20=#sum(tampella irtisanominen tavoite koota tieto #0(tampella konserni) kuuluva yhtiö suorittaa irtisanominen tieto lomautus lyhentää työviikko);

#q21=#sum(kera @ktm investointi matkailu tieto matkailuala yritys myöntää avustus laina investointi arvokas yhteenveto);

#q22=#sum(neste oy maakaasutoiminta yleiskuva neste maakaasutoiminto neste puuhailla maakaasu hankinta kenttä tuontisopimus jakelu verkosto rakentaminen markkinointi alue);

#q23=#sum(ydinvoimalaitos tuottaa radioaktiivinen jäte käsittely varastointi esimerkki ongelma riski sattua ydinjätevahinko);

#q24=#sum(aids levinneisyys #0(ey maa) vakava #0(aids tilanne) maa tieto esiintymämäärä kampanja tauti leviäminen ehkäistä toimi);

#q25=#sum(elintarvike tuontirajoitus poisto vaikutus suomi elintarviketeollisuus);

#q26=#sum(asuntotuotanto suhdanne suhdannevaihtelu suomi valtakunnallinen taso tilasto ennustetieto arvio rakentaminen asuntokauppa);

#q27=#sum(tieliikenne päästö suomi ulkomaa päästö kehittyä odottaa kehittyä lainsäädäntö vaikutus merkittävä katalysaattori yleistyminen vaikuttaa päästötaso katalysaattoritekniikka sinänsä kiinnostaa);

#q28=#sum(japani autoteollisuus investointi eurooppa tuotannollinen yhteistyö eurooppalainen autonvalmistaja maa japanilainen autotehdas suunnitella perustaa laajentaa tuotantomäärä trendi);

#q29=#sum(metsäteollisuus ympäristöinvestointi rajoittua vesiensuojelu liittyä investointi kemiallinen metsäteollisuus varsinainen puhdistamoinvestointi ympäristöystävällinen prosessi käyttöönotto);

#q30=#sum(kauppa aukioloaika selvittää vähittäiskauppa aukioloaika vapauttaminen koskea keskustelu kartoittaa kauppa järjestö ammattijärjestö kannanotto toimi);

#q31=#sum(pakkaus ympäristönsuojelukysymys kiinnostaa kulutustavarapakkaus kierrätysjärjestelmä kehittämiskokeilu kierrätys liittyä lainsäädäntö eri maa);

#q33=#sum(esko aho suomi #0(ey jäsenhakemus) aho suomi #0(ey jäsenyys) hakeminen liittyä mielipide kannanotto toimi arvio esko puhe toimi);

#q34=#sum(kauko juhantalo ydinvoimapuhe teko juhantalo perustelu 5 ydinvoimala juhantalo viedä ydinvoimalaratkaisu);

#q35=#sum(vihreä aloite välikysymys ehdotus puheenvuoro äänestyskäyttäytyminen suomi eduskunta tarkastelu ryhmä yksittäinen kansanedustaja);

Suomenkieliset karsitut kyselyt karsittuun hakemistoon

#q1=#sum(georg bushin mihail gorbatshov tapaamin helsing syysku 1990 neuvottelu käsitely asia tehdy päätöks sopimuks);

#q2=#sum(etel-amerik velkakr velkaantumisongelm kehittyny ongelm pyrity ratkaisem);

#q3=#sum(metsäteollisuud polkumyyntisyyt usa kiinnostav suomalaist paperinviej kohtalo polkumyyntisyytöst sisältö oikeudenkäyn tuloks);

#q4=#sum(jyväskyl kaupung maalaiskun kuntaliitoshank kartoit liitoshank kannattaj vastustaj mielipit perustelu arvio liitoks taloudellis vaikutuks porkkanarah);

#q6=#sum(varsov liito lakkauttamin muutosproses eri maide suhtautumis päätöks);

#q7=#sum(neuvostoliito lietua kohdistam taloussaarto kevä 1990 toim taloussaarto liityi näkyi lietua saaro lopettamis johtan tapahtum);

#q8=#sum(ira joukkotuhoas hävittämin ira persianlahd soda aseleposopimuks luovuttav kemiallis biologis ydinas tuotantotekniik yk vast ase inventoin hävittämis tehtäv suoritus onnistunu);

#q9=#sum(opec öljy hint tuotantomäär koskev päätöks);

#q10=#sum(president illiescu hallituks avu kaivosmiest väkivaltaisuud oppositio bukarest taustatieto tapahtum uhr jälkiselvittely);

#q11=#sum(namibia itsenäistymis liittyny yk rauhanturvaoperaatio tieto operaatio valmistelu liittyn tapahtum untag-jouko suomalaispataljoon toimin);

#q12=#sum(ey parlament asem yhteisö päätöksenteo selvit ey parlament asem suht komisio toimielim muutoks nykyis halutu halun demokraattin kontrol toimi ey);

#q13=#sum(carl bildt pohjoismain yhteistyö bildt pohjoism yhteistyö koskev lausuno erity bildt sanonu ruots suome yhteistyö);

#q14=#sum(jugoslavia presidenttineuvosto toimint koskev uutis tiedo istuno tehdy päätöks);

#q15=#sum(län itä-saks miehittäjävalto yhdysval iso-britan ransk neuvostoliito 2+4-neuvottelu sakso yhdistymis ratkaistav kysymyks riitakysymyks olen syntyn sopimuks);

#q17=#sum(valmet traktor kuljetusvälinetuotano kannattavuus kuljetusvälinetoimial kuuluv mets siirtokon kiskokalust transtech osakkuud henkilö kuorm-autoteollisuud jäte tarkastelu);

#q20=#sum(tampel irtisanomis tavoit koota tieto tamp-konsern kuuluv yhtiö suorittam irtisanomis tieto lomautuks lyhennety työviiko);

#q21=#sum(kera ktm investoin matkailu tieto matkailual yrityks myönnety avustuks laino investoint arvok yhteenvedo);

#q22=#sum(nest oy maakaasutoimin yleiskuv nest maakaasutoimino nest puuhailu maakaasu hankin kent tuontisopimuks jakelu verkosto rakentamin markkinoin alue);

#q23=#sum(ydinvoimalaitost tuottam radioaktiivist jät käsittely varastoint esimerk ongelm risk sattun ydinjätevahingo);

#q24=#sum(aids levinneisyys ey-mais vakav aids-tila mais tieto esiintymämäär kampanj taud leviäm ehkäisev toim);

#q25=#sum(elintarvik tuontirajoitust poisto vaikutus suome elintarviketeollisuut);

#q26=#sum(asuntotuotano suhdant suhdannevaihtelu suome valtakunnallin taso tilasto ennustetieto arvio rakentamis asuntokaup);

#q27=#sum(tieliikent päästö suome ulkom päästö kehittyn odot kehittyv lainsäädänö vaikutus merkittäv katalysaattor yleistymin vaikut päästötaso katalysaattoritekniik sinä kiino);

#q28=#sum(japan autoteollisuud investoin euroop tuotannollin yhteistyö eurooppalaist autonvalmistaj maih japanilais autoteht suunniteltu perustetu laajennetu tuotantomäär trend);

#q29=#sum(metsäteollisuud ympäristöinvestoin rajoitu vesiensuojelu liittyv investoint kemiallis metsäteollisuud varsinais puhdistamoinvestoin ympäristöystävällis proses käyttöönoto);

#q30=#sum(kaupa aukioloaj selvit vähittäiskaupo aukioloaiko vapauttam koskev keskustelu kartoit kaupa järjestöj ammattijärjestöj kannanoto toim);

#q31=#sum(pakkauks ympäristönsuojelukysymyks kiinnostav kulutustavarapakkaust kierrätysjärjestelm kehittämiskokeilu kierrätyks liittyv lainsäädäntö eri mais);

#q33=#sum(esko aho suome ey-jäsenhakemus aho suome ey-jäsenyyd hakemis liittyv mielipit kannanoto toime arvio esko puhe toim);

#q34=#sum(kauko juhantalo ydinvoimapuh teot juhantalo perustelu 5 ydinvoimal juhantalo vei ydinvoimalaratkaisu);

#q35=#sum(vihr aloit välikysymyks ehdotuks puheenvuoro äänestyskäyttäytymin suome eduskun tarkastelu ryhm yksittäis kansanedustaj);

Suomenkieliset ositetut perusmuotoiset kyselyt ositettuun perusmuotohakemistoon

#q1=#sum(george bush mihail gorbatshov tapaaminen helsinki #0(syys kuu) 1990 neuvottelu käsitelty asia tehty päätös sopimus);

#q2=#sum(#0(etelä amerikka) #0(velka kriisi) #0(velkaantumis ongelma) kehittyä ongelma pyrkiä ratkaista);

#q3=#sum(#0(metsä teollisuus) #0(polku myynti syyte) usa kiinnostava suomalainen #0(paperin viejä) kohtalo #0(polku myynti syytös) sisältö #0(oikeuden käynti) tulos);

#q4=#sum(#0(jyväs kylä) kaupunki #0(maalais kunta) #0(kunta liitos hanke) kartoittaa #0(liitos hanke) kannattaja vastustaja #0(mieli pide) perustelu arvio liitos taloudellinen vaikutus #0(porkkana raha));

#q6=#sum(#0(varsova liitto) lakkauttaminen #0(muutos prosessi) eri maa suhtautuminen päätös);

#q7=#sum(#0(neuvosto liitto) liettua kohdistaa #0(talous saarto) kevät 1990 toimi #0(talous saarto) liittyä näkyä liettua saarto lopettaminen johtaa tapahtuma);

#q8=#sum(irak #0(joukko tuho ase) hävittäminen irak #0(persian lahti) sota #0(ase lepo sopimus) luopua kemiallinen biologinen #0(ydin ase) #0(tuotanto tekniikka) yk vastata ase inventointi hävittäminen tehtävä suoritus onnistua);

#q9=#sum(opec öljy hinta #0(tuotanto määrä) koskea päätös);

#q10=#sum(presidentti @illiescun hallitus apu #0(kaivos mies) #0(väki valtaisuus) oppositio bukarest #0(tausta tieto) tapahtuma uhri #0(jälki selvittely));

#q11=#sum(namibia itsenäistyminen liittyä yk #0(rauhan turva operaatio) tieto operaatio valmistelu liittyä tapahtuma #0(@untag joukko) #0(suomalais pataljoona) toiminta);

#q12=#sum(ey parlamentti asema yhteisö #0(päätöksen teko) selvittää ey parlamentti asema suhde komissio #0(toimi elin) muutos nykyinen haluttu halunnut demokraattinen kontrolli toimia ey);

#q13=#sum(carl @bildt #0(pohjois mainen) #0(yhteis työ) @bildtin #0(pohjois mainen) #0(yhteis työ) koskea lausunto eritty @bildt sanoa ruotsi suomi #0(yhteis työ));

#q14=#sum(jugoslavia #0(presidentti neuvosto) toiminta koskea uutinen tieto istunto tehty päätös);

#q15=#sum(länsi #0(itä saksa) #0(miehittäjä valta) #0(yhdys valta) #0(iso britannia) ranska #0(neuvosto liitto) #1(2+4 neuvottelu) saksa yhdistyminen ratkaista kysymys #0(riita kysymys) olennainen syntyä sopimus);

#q17=#sum(valmet traktori #0(kuljetus väline tuotanto) kannattavuus #0(kuljetus väline toimi ala) kuuluva metsä #0(siirto kone) #0(kisko kaluste) @transtech osakkuus henkilö #0(kuorma auto teollisuus) jättää tarkastelu);

#q20=#sum(tampella irtisanominen tavoite koota tieto #0(tampella konserni) kuuluva yhtiö suorittaa irtisanominen tieto lomautus lyhentää #0(työ viikko));

#q21=#sum(kera @ktm investointi matkailu tieto #0(matkailu ala) yritys myöntää avustus laina investointi arvokas #0(yhteen veto));

#q22=#sum(neste oy #0(maa kaasu toiminta) #0(yleis kuva) neste #0(maa kaasu toiminto) neste puuhailla #0(maa kaasu) hankinta kenttä #0(tuonti sopimus) jakelu verkosto rakentaminen markkinointi alue);

#q23=#sum(#0(ydin voima laitos) tuottaa #0(radio aktiivinen) jäte käsittely varastointi esimerkki ongelma riski sattua #0(ydin jäte vahinko));

#q24=#sum(aids levinneisyys #0(ey maa) vakava #0(aids tilanne) maa tieto #0(esiintymä määrä) kampanja tauti leviäminen ehkäistä toimi);

#q25=#sum(#0(elin tarvike) #0(tuonti rajoitus) poisto vaikutus suomi #0(elin tarvike teollisuus));

#q26=#sum(#0(asunto tuotanto) suhdanne #0(suhdanne vaihtelu) suomi #0(valta kunnallinen) taso tilasto #0(ennuste tieto) arvio rakentaminen #0(asunto kauppa));

#q27=#sum(#0(tie liikenne) päästö suomi #0(ulko maa) päästö kehittyä odottaa kehittyä #0(lain säädäntö) vaikutus merkittävä katalysaattori yleistyminen vaikuttaa #0(päästö taso) #0(katalysaattori tekniikka) sinänsä kiinnostaa);

#q28=#sum(japani #0(auto teollisuus) investointi eurooppa tuotannollinen #0(yhteis työ) eurooppalainen #0(auton valmistaja) maa japanilainen #0(auto tehdas) suunnitella perustaa laajentaa #0(tuotanto määrä) trendi);

#q29=#sum(#0(metsä teollisuus) #0(ympäristö investointi) rajoittua #0(vesien suojelu) liittyä investointi kemiallinen #0(metsä teollisuus) varsinainen #0(puhdistamo investointi) #0(ympäristö ystävällinen) prosessi #0(käyttöön otto));

#q30=#sum(kauppa #0(auki olo aika) selvittää #0(vähittäis kauppa) #0(auki olo aika) vapauttaminen koskea keskustelu kartoittaa kauppa (ammatti järjestö) #0(kannan otto) toimi);

#q31=#sum(pakkaus #0(ympäristön suojelu kysymys) kiinnostaa #0(kulutus tavara pakkaus) #0(kierrätys järjestelmä) #0(kehittämis kokeilu) kierrätys liittyä #0(lain säädäntö) eri maa);

#q33=#sum(esko aho suomi #0(ey jäsen hakemus) aho suomi #0(ey jäsenyys) hakeminen liittyä #0(mieli pide) #0(kannan otto) toimi arvio esko puhe toimi);

#q34=#sum(kauko juhantalo #0(ydin voima puhe) teko juhantalo perustelu 5 #0(ydin voimala) juhantalo viedä #0(ydin voimala ratkaisu));

#q35=#sum(vihreä aloite #0(väli kysymys) ehdotus #0(puheen vuoro) #0(äänestys käyttäytyminen) suomi #0(edus kunta) tarkastelu ryhmä yksittäinen #0(kansan edustaja));