Sari Kaitaniemi

Kyselyn automaattinen laajentaminen synonyymeilla

Sari Kaitaniemi: Kyselyn automaattinen laajentaminen synonyymeillä. [Automatic Query Expansion by using synonyms] Informaatiotutkimus 22(4), 105-120.

The objective of this study was to experiment automatic query expansion with synonyms using a probabilistic information retrieval system, InQuery. The document collection consists of 54 000 Finnish newspaper articles from three newspapers. Two sources of additional terms were compared: a general synonym thesaurus and a thesaurus built specifically for the document collection used in the experiment. Also two query structures were compared: a query type with no operators used to formulate the query and a structured query type, where the concept structure was clearly built with operators. In the results the only expansion, that showed any improvement to the original query was the structured query expanded with the thesaurus specifically built for the document collection.

Address: Sari Kaitaniemi, Department of Information Studies, FIN-33014 University of Tampere, Finland, email sari.kaitaniemi@uta.fi

1. Johdanto

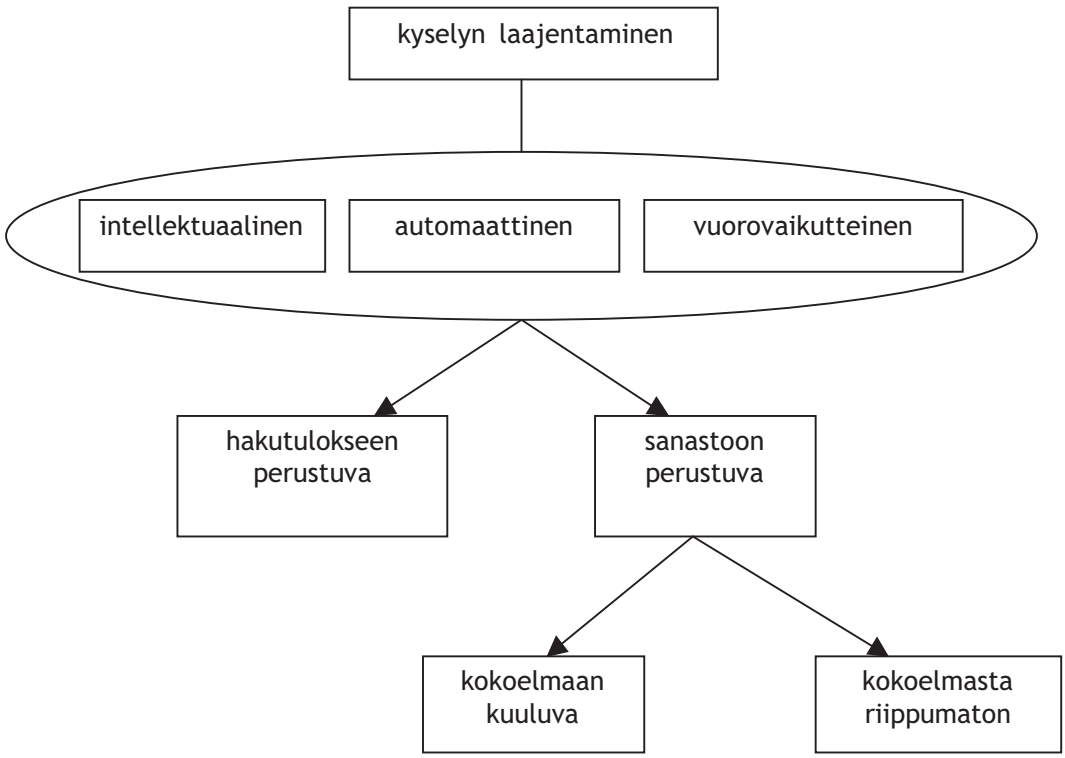
Luonnollinen kieli rikastuttaa ihmisten välistä kommunikointia - ja aiheuttaa ongelmia tekstitiedonhakuun. Tietokoneella tehtävä tiedonhaku on merkkijonoihin perustuva, eksakti tapahtuma. Yhden kirjaimen ero sanan kirjoitusasussa saattaa aiheuttaa sen, että arvokas dokumentti jää löytymättä. Luonnollisessa kielessä samaan asiaan voidaan viitata useilla eri nimillä, synonyymeilla. Lisäksi asiat voidaan ilmaista joko yksityiskohtaisesti tai yleisellä tasolla. Pro gradu -tutkielmassani tutkin tiedonhakua suomenkielisestä artikkelitietokannasta. Tutkimus oli luonteeltaan empiirinen laboratoriotutkimus. Tutkimusongelma on: miten suomenkielisen kyselyn automaattinen laajentaminen synonyymeilla vaikuttaa hakutulokseen probabilistisessa hakujärjestelmässä.

Tietoa hakee yhä useammin tiedon tarvitsija itse, kun kaupalliset ja Internetin tietojärjestelmät tuovat tiedon tarvitsijoiden ulottuville. Kokematon tiedonhakija tekee usein lyhyitä kyselyitä eikä välttämättä hallitse kyselynmuodostuksen tekniikoita.

Äkkiseltään vuorovaikutteinen kyselyn laajentaminen voisi tuntua hyvältä ratkaisulta; tarjotaan hakijalle laajennusavaimia sanastosta, ja annetaan hänen itse valita mielestään hyödylliset. Magennis & Rijsbergenin (1997, 74-81) mukaan kokemattomat käyttäjät eivät yleensä osaa valita hyödyllisiä laajennusavaimia ja hakutulos useimmin ei ainakaan parane vuorovaikutteisesti laajentamalla. Satunnaista tiedonhakijaa auttaisi todennäköisesti parhaiten järjestelmä, joka osaa itse laajentaa kyselyä tarkoituksenmukaisesti.

2. Kyselyn laajentaminen

Kyselyn laajentamista on tutkittu paljon erilaisin menetelmin. Efthimiadisin (1996, 122) mukaan kyselyn laajentamisesta (Query Expansion) on kyse, kun alkuperäistä kyselyä täydennetään uusilla avaimilla hakutuloksen parantamiseksi. Kysely voidaan laajentaa intellektuaalisesti, automaattisesti tai vuorovaikutteisesti. Laajennusavaimet voidaan ottaa joko hakutuloksen dokumenteista, jolloin laajennusavaimet perustuvat relevanssi-



Kuvio 1: Kyselyn laajentaminen: menetelmät ja termien lähteet (Efthimiadis 1996, 124)

palautteeseen, tai hakuprosessin ulkopuolisesta sanastosta, joka voi olla joko dokumenttikokoel maan kuuluva tai siitä täysin riippumaton. Nämä vaihtoehdot on esitetty kuviossa 1. (Efthimiadis 1996, 122-124.)

Kristensen (1992) tutki hakutesauruksen käyttöä kyselyn laajentamisessa. Tutkimus tehtiin suomenkielisellä Aamulehden noin 225 000 artikkelia sisältävässä tietokannalla Boolen logiikkaan perustuvalla hakujärjestelmällä rakenteisilla kyselyillä. Kysely on rakenteinen,

kun siinä esiintyvien hakutermien väliset suhteet on ilmaistu operaattorein, esim. Boolen operaattorein. Tyypillinen rakenteinen kysely toteuttaa lohkostrategiaa: samaa hakukäsitettä edustavat eli keskenään vaihtoehtoiset hakutermit on yhdistetty disjunktiolla eli OR-operaattorilla ja näin muodostuneet termifasetit yhdistetään konjunktiolla eri AND-operaattorilla. Jos kyselyssä ei ole käytetty operaattoreita määrittämään termien välisiä suhteita, kysely on rakenteeton eli litteä.

Kristensenin tutkimuksessa hakuaiheita oli 30. Perushakuja laajennettiin 1) synonyymeillä, 2) suppeammilla termeillä, 3) rinnakkaistermeillä ja 4) kaikilla edellisillä ryhmillä (=laajin haku). Laajin haku kaksinkertaisti perushaun tuloksen suhteellisen saannin ($p<0,0001$). Tarkkuus taas heikkeni noin kymmenen prosenttiyksikköä ($p<0,01$). Synonyymeilla, suppeammilla termeillä ja rinnakkaistermeillä laajentamisen saannit ja tarkkuudet olivat varsin samanlaisia, mutta tulosjoukkojen artikkeleissa oli vähän samoja. Paras saanti saatiin laajimmalla haulla.

Voorhees (1994) tutki kyselyn laajentamista intellektuaalisesti laajalla yleistesauruksella, WordNetillä. Testikokoelmana käytettiin TREC-kokoelmaa, joka sisälsi 742 000 englanninkielistä dokumenttia sanomalehdistä, teknisten kirjoitusten abstrakteista ja Federal Registeristä. Kyselyjä tehtiin 50. Testin hakujärjestelmä oli vektorimalliin perustuva SMART. (Voorhees, 1994.) Voorheesin kyselyt olivat vektoreita, jotka koostuivat eri käsitetyyppejä edustavista alavektoreista. Kysely laajennettiin lisäämällä synonyymit

kyselyvektoriin. Kyselyavaimet painotettiin siten, että alkuperäisen kyselyavaimen paino oli aina 1 ja lisättyjen avainten paino vaihteli välillä 0,1-2. Kokeessa laajennusavaimiksi valittiin kaikki avainryhmät, jotka liittyivät sanastossa suoraan kyselyavaimeen tasoilla synonyymit, alemmat ja ylemmät termit. Voorheesin kokeessa laajentamisen vaikutus hakutulokseen oli selvästi heikko. Perusyselyjä oli kolmen pituisia: 52,54 sanaa, 29,22 sanaa ja 11,02 sanaa. Lyhin testattu kyselytyyppi oli ainoa, jossa laajentaminen paransi hakutulosta merkitsevästi: se parani 35 % mitattuna tarkkuuksien keskiarvolla yli 11 pisteen saantitason (0,0…1,0). (Voorhees 1994.)

Kekäläinen (1999) tutki fasettipohjaisen kyselyn kompleksisuuden, laajentamisen ja rakenteen vaikutuksia hakutulokseen. Tutkimukset suoritettiin probabilistisella InQuery-järjestelmällä suomenkielisessä TUTK-tietokannassa (kuvaillaan myöhemmin). Kyselyt tehtiin intellektuaalisesti käsitefaseteista ja laajentaminen tapahtui automaattisesti ExpansionToolin avulla (ks. Järvelin, Kekäläinen, Niemi, 2001). Kekäläisen paras tulos saavutettiin laajentamalla kyselyjä yhdessä synonyymeilla, suppeammilla, laa-jemmilla ja rinnakkaisavaimilla käyttäen käsiterakenteeseen perustuvaa rakenteista kyselytyyppiä.

3. Tutkimusasetelma

3.1 Testitietokanta

Esiteltävän kokeen tietokanta on TUTK, joka sisältää noin 54 000 artikkelia suomalaisista sanomalehdistä. Artikkeleista noin 24 500 on peräisin Aamulehden ulkomaanosastolta, noin 16 800 Keskisuomalaisen eri osastoilta ja noin 14 000 artikkelia Kauppalehdestä. (Sormunen 1993, 64.) Kokoelmassa on 35 hakuaihetta. Kekäläinen (1999, 58-59) karsi tutkimuksessaan aiheet 30:een aiheen laajennettavuuden perusteella: pois jätettiin aiheet, joissa laajentaminen ei ollut mahdollista (esim. hakutehtävä sisälsi vain tai pääasiassa erisnimiä). Omassa työssäni käytän näitä 30 aihetta. Aiheet on lueteltu liitteessä 1.

TUTK-tietokannan (jatkossa TUTKin) dokumenttien relevanssi on arvioitu neliportaisella asteikolla (Sormunen 1993, 72; tässä Kekäläinen 1999, 96).

o Relevanssitasolla 0 dokumentti ei sisällä lainkaan aiheeseen liittyvää informaatiota, se ei siis sisälly aiheen saantikantaan.

o Tasolla 1 Dokumentti sisältää vain viittauksen aiheeseen, yhden lauseen tai faktan.

o Tasolla 2 dokumentti sisältää jossain määrin aiheeseen liittyvää informaatiota. Jos aihe on dokumentin pääteema, sitä on käsitelty lyhyesti tai pinnallisesti, tai aihe on dokumentin sivuteema. Aihetta on käsitelty noin yhden kappaleen verran.

o Tasolla 3 aihe on dokumentin pääteema ja informaatiosisältö on merkittävä. Laajuus on vähintään kaksi kappaletta, neljä lausetta tai faktaa. (Sormunen 1994, 71-72.)

Tekemässäni tutkimuksessa hakutuloksia arvioitiin kolmella erilaisella saantikannalla. Tasolla kaikki relevantit relevanssikorpuksessa ovat mukana dokumentit tasoilta 1-3. Tasolla relevantit relevanssikorpukseen kuuluvat dokumentit tasoilta 2 ja 3. Tasolla erittäin relevantit tulosjoukkoon vain tason 3 dokumentit.

TUTKin saantikantojen koko eli kunkin hakuaiheen relevanttien dokumenttien lukumäärä vaihtelee aineistossa suuresti. Esimerkiksi kaikki relevantit -tasoisia dokumentteja aiheeseen 10 (untag) on 143 ja aiheeseen 5 (varso) 129. Erittäin relevantit -tasoisia dokumentteja on aiheisiin 12 (bildt) ja 16 (tampel) vain yksi. Relevanttien dokumenttien vähyys vaikuttaa selvästi esimerkiksi keskiarvotarkkuuksiin. Kun relevantteja dokumentteja on vain muutama, yhden löytyminen tai löytymättä jääminen tekee kymmenien prosenttien eron ko. hakutuloksen tarkkuuteen.

3.2 Hakujärjestelmä

Kokeen hakujärjestelmä, InQuery, on todennäköisyyslaskentaan perustuva (probabilistinen), osittaistäsmäyttävä tiedonhakujärjestelmä. (Callan, Croft & Harding, 1992.) InQuery perustuu probabilistiseen hakumalliin, päättelyverkkoon. Haku toimii siten, että dokumentin esitysmuotoa verrataan kyselyn esitysmuotoon niiden sisältämien sanojen tilastollisten ominaisuuksien perusteella. Dokumentin esitysmuoto voi olla esimerkiksi sanoja, fraaseja, tekstikappaleita tai manuaalisesti annettuja avainsanoja. Kysely voi olla joko luonnollista kieltä tai operaattorein muodostettu hakulause. (Broglio, Callan & Croft, 1994).

InQueryllä hakijalla on käytössään laaja joukko operaattoreita. Tutkimuksessani käytin seuraavia:

Sum: sum-solmun arvo on sen kattamien avainten painojen keskiarvo.
Syn taas käsittelee sisältämiään avaimia tai fasetteja saman avaimen esiintyminä. (Applied Computing Systems Institute of Massachusetts = ACSIOM, Inc.,1996.) Syn-solmun arvo lasketaan seuraavalla kaavalla:

$$0{,}4 + 0{,}6 * \left(\frac{\sum_{i \in s} tf_{ij}}{\sum_{i \in s} tf_{ij} + 0{,}5 + 1{,}5 * \frac{dl_j}{adl}} \right) * \left(\frac{\log\left(\frac{N + 0{,}5}{df_s}\right)}{\log(N + 1{,}0)} \right)$$

missä

tf_{ij} = avaimen i frekvenssi dokumentissa j
S = syn-operaattorin yhdistämä hakuavainjoukko
dl_j = dokumentin j avainten määrä
adl = kokoelman dokumentin keskiarvopituus
N = kokoelman dokumenttien lukumäärä
df_s = niiden dokumenttien lukumäärä, jotka sisältävät vähintään yhden joukon S avaimen. (Kekäläinen 1999, 28.)

Operaattori **uwn** on läheisyysoperaattori, unordered window n. Se edellyttää kaikkien hakuavaimien esiintyvän n:n sanan kokoisessa ikkunassa vapaassa järjestyksessä (Applied Computing Systems Institute of Massachusetts, Inc.,1996). Esimerkkejä operaattorien käytöstä löytyy luvusta 3.3.

3.3 Kyselyjen laajentamistavat

Kokeissani laajensin kyselyt kahdella sanastolla: kaupallisella, yleisluontoisella Finthes-synonyymisanastolla sekä Kekäläisen (1999, 59) väitöskirjatutkimuksessaan TUTKiin räätälöimällä hierarkkisella käsitetesauruksella. Kutsun sitä tässä työssä tesaurukseksi. Tämä tesaurus koostuu käsitteistä, ilmauksista ja täsmäytysmalleista ja niiden välisistä suhteista. Tesauruksen käsitteiden lukumäärä on 832 ja niiden ilmausten määrä on 1345. Käsitteiden väliset suhteet ovat joko rinnakkaistermi- tai hierarkkisia suhteita. Synonyymisuhteita ei esiinny käsitteiden, vaan käsitteen ilmausten, välillä. (Kekäläinen 1999, 59-67.) Tämän tutkimuksen kyselyjä laajennettiin vain synonyymeilla.

Finthes on Lingsoftin tuote. Finthesissä on noin 7400 synonyymiryhmää ja niissä yhteensä noin 26 300 synonyymia. Sama sana voi olla useammankin sanan synonyymina, eri synonyymeja on noin 21 700. Synonyymeja on siis keskimäärin 3,55 kussakin ryhmässä. (Ronkainen, 2002.)

Vertasin viidenlaisia kyselyjä:

- v peruskysely
- Ø laajentamaton litteä kysely
- v litteä Finthes-kysely
- Ø laajennettu Finthesillä
- v litteä tesauruskysely
- Ø laajennettu TUTK-tesauruksen synonyymeilla
- v rakenteinen Finthes-kysely
- Ø rakenteinen, laajennettu Finthesillä
- v rakenteinen tesauruskysely
- Ø rakenteinen, laajennettu TUTK-tesauruksen synonyymeilla

Mallinsin laajennosten tekemisessä automaattista laajentamista. Tein Finthes-laajennokset seuraavalla periaatteella:

1. Syötin hakuavaimen Finthesiin.
2. Jos Finthes löysi avaimelle synonyymeja, syötin synonyymit Fintwoliin.
3.a. Jos Fintwol hyväksyi synonyymin sellaisenaan, hyväksyin sen laajennusavaimeksi.
b. Jos Fintwol antoi syötetylle synonyymille muun perusmuodon, hyväksyin sen laajennusavaimeksi

Finthes käsittelee annetun sanan kaikki mahdolliset tulkintavaihtoehdot ja antaa synonyymit samassa taivutusmuodossa, kuin missä analysoi annetun sanan olevan. Fintwol puolestaan analysoi syötettyjen sanojen mahdolliset kantasanat, ja antaa mahdolliset kantasanat perusmuodossa. Hyväksyin laajennusavaimiksi nekin Finthes – Fintwol -tuotokset, joista ihminen osaa heti sanoa semanttisen tietämyksensä perusteella, etteivät ne ole hakuavaimen synonyymeja ainakaan tässä kontekstissa. Yksinkertainen automaattinen kyselynlaajennin ei osaa karsia laajennusavainehdokkaita semantiikan perusteella.

Esimerkiksi hakuavain Suomi (aihe 21 elint) saa Finthesistä synonyymeikseen piiskaa, ruoski, vitso ja piiskasi, ruoski, vitsoi. Fintwolilla perusmuotoistamalla Suomi-avaimen synonyymeiksi Finthes-laajennoksiin tuli piiskata, ruoskia,

vitsoa. Vaikka tällainen laajentaminen lisää hälyä hakutulokseen, on sitä yksinkertaisin keinoin mahdotonta välttää automaattisessa laajentamisessa. Useimmissa tapauksissa Finthes antoi enimmäkseen hyödyllisiä avaimia laajennettuun kyselyyn.

Tesaurus on Kekäläisen (1999, 59) väitöski rjatutkimuksessaan TUTKia varten rakentama hierarkkinen, käsitepohjainen tesaurus. Tesauruksen käsitteet voivat muodostua useammasta sanasta, esimerkiksi kemiallinen metsäteollisuus. Tämän käsitteen synonyymiksi tesaurus antaa kemiallinen puunjalostusteollisuus. TUTKissa kaikki sanat on perusmuotoistettu. Tämän vuoksi tesauruksen sanat annetaan perusmuotoisina. Esimerkiksi vesiensuojelu saa synonyymeikseen

- v suojella vesi
- v vesi suojeleminen
- v vesistö suojelu
- v suojella vesistö
- v vesistö suojeleminen.

Rakenteiset kyselyt muodostin käyttämällä syn-operaattoria synonyymit kokoavana operaattorina ja yhdistämällä synonyymifasetit sum-operaattorilla.

Esimerkiksi kysely 19 (ydinv) peruskyselynä:

#q19 = #sum(ydinvoimala ydinjäte käsittely varastointi onnettomuus ongelma);

Litteänä Finthes-laajennoksena:

#q19 = #sum(ydinvoimala ydinjäte käsittely työstö muokkaus työstäminen manipulointi manipulaatio ruodinta pohdinta tarkastelu varastointi talteenpano tallennus talteenotto talletus säilytys pito tallessapito onnettomuus tapaturma turma vahinko haaveri ongelma kysymys asia juttu probleema seikka pulma probleemi pähkinä tehtävä);

Rakenteisena Finthes-laajennoksena:

```
#q19 = #sum(ydinvoimala ydinjäte
            #syn(käsittely työstö muokkaus
            työstäminen  manipulointi manipu
            laatio ruodinta pohdinta tarkastelu)
               #syn(varastointi talteenpano
                    tallennus talteenotto
            talletus  säilytys pito tallessapito)
            #syn(onnettomuus tapaturma
                 turma vahinko haaveri)
            #syn(ongelma kysymys
                 asia  juttu probleema
                 seikka pulma probleemi
                 pähkinä tehtävä));
```

Litteänä tesauruslaajennoksena:

#q19 = #sum(ydinvoimala ydinvoimalaitos atomivoimala atomivoimalaitos ydinjäte #uw3(radioaktiivinen jäte)ydinvoimajäte ydinvoimalajäte käsittely käsitteleminen käsitellä varastointi varastoiminen varastoida säilytys säilyttäminen säilyttää taltiointi taltioiminen taltioida onnettomuus tapaturma vahinko turma vaurio haaveri ongelma pulma probleema ongelmallinen pulmallinen problemaattinen);

Rakenteisena tesauruslaajennoksena:

```
#q19 = #sum(#syn(ydinvoimala ydinvoima
                 laitos atomivoimala atomi
                 voimalaitos)
          #syn(ydinjäte #uw3(radioaktiivinen
                 jäte)
                  ydinvoimajäte ydinvoimalajäte)
          #syn(käsittely käsitteleminen käsitellä)
          #syn(varastointi varastoiminen varastoi-
                 da  säilytys säilyttäminen säilyttää
                 taltiointi taltioiminen taltioida)
        #syn(onnettomuus tapaturma vahinko
                 turma vaurio haaveri)
         #syn(ongelma pulma probleema
            ongelmallinen  pulmallinen
                 problemaattinen));
```

Sanaliitot muodostin peruskyselyissä läheisyys-operaattorilla uwn. N:ksi, eli ikkunan kooksi, asetin liiton osien lukumäärän pyöristettynä seuraavaan parittomaan lukuun. Finthes-laajennoksissa sanaliitot laajennettiin osa kerrallaan. Sanaliittojen osillekin löytyi synonyymeja. Laajennetut sanaliitot muodostin samalla tavalla eli yhdistin osat läheisyysoperaattorilla uwn. Sanaliiton synonyymifasetit yhdistin syn-operaattorilla. Esimerkiksi kyselyssä 11 (eyval) esiintyi sanaliitto EY:n parlamentti. Perusmuotoistettuna tämä muuntui sanapariksi EY parlamentti. Koska InQuery käsittelee kaikki sanat pienellä alkukirjaimella kirjoitettuna, InQueryn hakuavaimeksi tuli #uw3(ey parlamentti).

Peruskysely:

#syn(... #uw3(ey parlamentti)...);

Laajennettu (Finthes):

#syn(... #uw3(ey #syn(parlamentti kansanedus tuslaitos eduskunta))...);

T

ekemässäni pikatestissä tämä menetelmä osoittautui paremmaksi kuin sanaliiton kaikkien osien yhdistäminen tasa-arvoisiksi syn-operaattorilla. 15 kyselyllä menetelmien välinen ero osoittautui tilastollisesti melko merkitseväksi ja käytännössä kiinnostavaksi. Menetelmä selitetään tarkemmin pro gradu –työssäni (2002).

Laajentamisen vaikutusta hakutulokseen mittasin ensinnäkin saannin ja tarkkuuden avulla. Ne ovat nykyään yleisimmät tiedonhaun tehokkuuden mittarit (Ks. esim. Alaterä & Halttunen 2002; Järvelin 1995; Salton & McGill, 1983). Tässä työssä hakumenetelmien eroja tarkastellaan 10 saantipisteen keskiarvotarkkuuksien avulla. Kullekin kyselylle lasketaan sen tarkkuusluku saantipisteissä 0,1…1,0. Näistä lasketaan keskiarvot yli eri tehtäviä edustavien kyselyjen saantitarkkuuskäyrien piirtämiseksi ja vielä edellisistä yli saantipisteiden hakumenetelmäkohtaisen vertailuluvun saamiseksi. Koska minua kiinnosti nimenomaan menetelmän käytännön merkitys, arvioin tuloksia myös Karen Sparck Jonesin (1974) ”peukalosäännöllä”. Hänen mielestään alle viiden prosenttiyksikön ero menetelmien välillä ei ole huomion arvoinen, 5-10 prosenttiyksikön ero on kiinnostava ja vasta yli 10 prosenttiyksikön ero on huomattava.

Kolmas mittari, jolla tutkin laajentamismenetelmien välisiä eroja, on kumuloitu hyöty. Käyttäjän kannalta olisi hyödyllistä, jos relevanteimmat dokumentit löytyisivät tuloslistan alkupäästä. Harva tiedon tarvitsija jaksaa selata muutamaa kymmentä viitettä tai dokumenttia enempää. Jos relevanssiarvio on binäärinen, relevantteihin dokumentteihin lukeutuu niin erittäin kuin marginaalisestikin relevantteja dokumentteja. Järjestelmän kykyä löytää erittäin relevantit dokumentit voidaan arvioida, kun dokumenttien relevanssi on arvioitu monitasoisesti ja eri relevanssitasojen hakutulosta verrataan keskenään. Järjestelmän kykyä saada relevanteimmat dokumentit tulosjoukon kärkeen mittaa kumuloitu hyöty (cumulated gain, CG). (Järvelin & Kekäläinen, 2002.) Kumuloitu hyöty lasketaan tuloslistassa olevan dokumentin järjestysluvun ja relevanssiarvon tunnusluvun perusteella. Tuloslistassa dokumentin järjestysluku korvataan sen relevanssiarvolla. Kunkin dokumentin kohdalla näkyy siihen mennessä kertynyt hyöty, joka on dokumentin ja sitä edeltävien dokumenttien relevanssiarvojen summa.

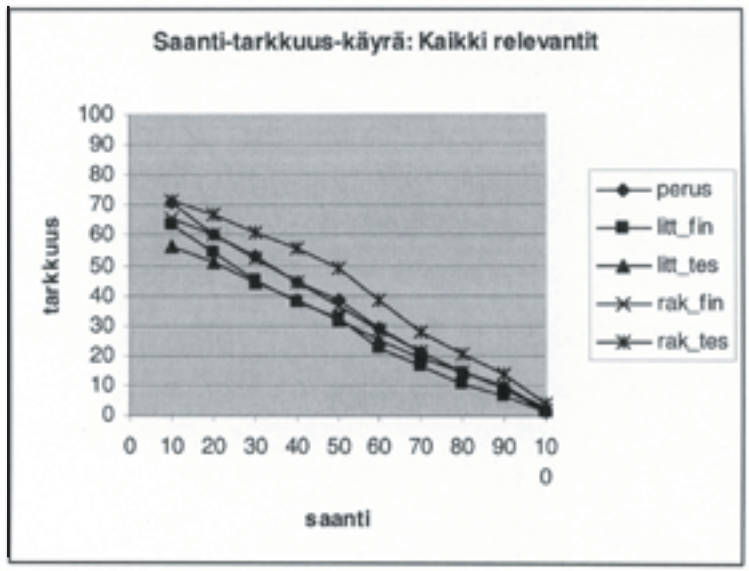
Tulosten tilastollista merkitsevyyttä tarkastelen Friedmanin kaksisuuntaisella järjestyslukutestillä (ks. esim. Siegel, 1989). Friedmanin testi on ei-parametrinen testi eli sitä käytetään, kun otokset eivät noudata normaalijakaumaa, mikä on yleensä tilanne tiedonhaun tutkimuksessa (Kekäläinen 1999, 98-99). Friedmanin testiä suositellaan käytettäväksi, kun vertailtavana on enemmän kuin kaksi toisistaan riippuvaa otosta. Kekäläinen (1999, 101) käytti tutkimuksessaan Conoverin versiota Friedmanin testistä, koska se on herkempi osoittamaan merkitsevän eron kuin Siegelin (1989) versio. Omissa kokeissani käytän samoin Conoverin versiota.

4. Tutkimustulokset

4.1 Kaikki relevantit

Kaikkien relevanttien dokumenttien relevanssikorpuksessa tehtyjen hakujen saantitarkkuustulokset ovat kuviossa 2 ja taulukossa 1. Niistä ilmenee, että rakenteinen tesauruslaajennos tuotti kaikilla saantiarvoilla paremman hakutuloksen kuin mikään muu kyselytyyppi. Lisäksi se oli ainoa laajennusmenetelmä, joka oli kaikilla tasoilla parempi kuin peruskysely. Kaikkien muiden laajennusmenetelmien tulos oli siis heikompi kuin peruskyselyn. Huonoin tarkkuus vaihtelee alhaisilla saantitasoilla litteiden laajennosten välillä. Korkeilla saantitasoilla ja 11 tason keskiarvotarkkuuden perusteella litteä Finthes-laajennos on heikoin kyselytyyppi.

Hakumenetelmien väliset erot ovat erittäin merkitseviä eroja tässä relevanssikorpuksessa. Jopa Friedmanin testi antoi tunnusluvuksi 0,000000000, eli tilastotestiohjelman lasku-tarkkuus ei riittänyt erojen merkitsevyyden suuruuden kuvaamiseen. Litteä laajentaminen on siis selvästi epäedullinen laajentamismenetelmä, ja rakenteinen tesauruslaajennos selvästi edullinen menetelmä. Molemmat rakenteiset menetelmät ovat tilastollisesti erittäin merkitsevästi ($p<0,001$, taulukko 2) molempia



Kuvio 2: Kaikki relevantit dokumentit - saanti—tarkkuus –käyrä eri kyselymenetelmillä

Taulukko 1: Kaikki relevantit dokumentit - tarkkuus saantitasoittain eri kyselymenetelmillä (paras tarkkuus varjostettu, huonoin tarkkuus alleviivattu

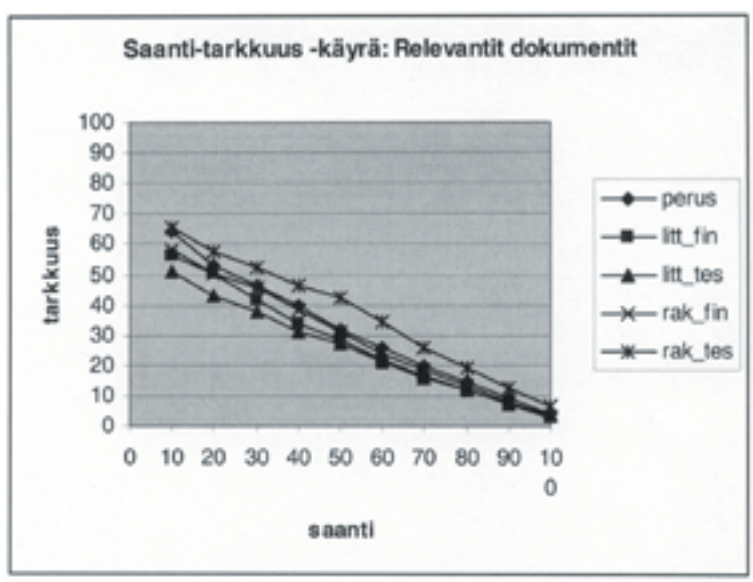
saanti	perus	litt_fin	litt_tes	rak_fin	rak_tes
10	70,8	63,7	56,6	65,3	71,5
20	60,4	54,1	51,2	60,0	67,0
30	53,0	45,0	44,6	52,0	60,6
40	44,6	38,0	38,6	44,2	55,6
50	38,3	32,2	32,1	36,5	48,8
60	29,1	22,4	25,1	28,5	38,3
70	21,0	16,3	19,1	21,0	28,0
80	14,7	10,8	14,4	14,7	20,7
90	9,1	6,8	9,6	10,1	14,1
100	1,3	1,2	2,9	1,8	4,0
keskiarvo	34,2	29,0	29,4	33,4	40,9

Taulukko 2: Friedmanin testi. Kaikki relevantit

	perus	litt_fin	litt_tes	rak_fin
litt_fin	***			
litt_tes	**	-		
rak_fin	-	***	***	
rak_tes	**	***	***	*

Taulukossa 2
- = ei merkitsevää eroa
* = p<0,05 melko merkitsevä ero
** = p<0,005 varsin merkitsevä ero
*** = p<0,001 erittäin merkitsevä ero
tummennettu ne ruudut, joissa vaakarivin menetelmä on parempi kuin pystyrivin menetelmä.

Kuvio 3: Relevantit dokumentit - saanti-tarkuus käytä eri kyselymentelmillä



litteitä menetelmiä parempia. Sparck Jonesin peukalosäännön perusteella ainoa käytännössä merkittävä ero on rakenteisen tesauruslaajennuksen ja molempien litteiden laajennusten välillä (11,9 ja 11,5 prosenttiyksikköä). Rakenteisen tesauruslaajennoksen keskiarvotarkkuus on ainoa peruskyselyä parempi keskiarvotarkkuus, ero on 6,7 prosenttiyksikköä. Tämä ero on tilastollisesti varsin merkitsevä (p<0,005), mutta Sparck Jonesin mukaan tämä ero on vain kiinnostava, ei käytännössä tärkeä. TUTK-tesauruksella rakenteinen laajennus on kokeen paras menetelmä ja samalla sanastolla litteä laajennus on toisiksi huonoin menetelmä.

Taulukko 3: Relevantit dokumentit - tarkkuus saantitasoittain eri kyselymenetelmillä

saanti	perus	litt_fin	litt_tes	rak_fin	rak_tes
10	64,3	56,4	50,9	58,4	65,6
20	53,3	50,1	43,3	50,1	57,7
30	46,2	42	37,5	45,6	52,0
40	39,5	34,1	31,3	38,7	46,5
50	32,0	28,2	26,9	31,4	42,2
60	26,1	21,8	21,4	23,9	34,4
70	19,8	16,3	16,2	18,4	25,6
80	14,6	11,7	12,2	13,0	19,1
90	9,5	7,2	7,6	8,2	12,7
100	4,2	2,7	4,3	3,3	6,4
keskiarvo	30,9	27,1	25,1	29,1	36,2

Taulukko 4: Friedman testi. relevantit

	perus	litt_fin	litt_tes	rak_fin
litt_fin	**			
litt_tes	**	-		
rak_fin	-	**	**	
rak_tes	*	***	***	**

Taulukossa 4
- = ei merkitsevää eroa
* = p<0,05 melko merkitsevä ero
** = p<0,005 varsin merkitsevä ero
*** = p<0,001 erittäin merkitsevä ero
tummennettu ne ruudut, joissa vaakarivin menetelm on parempi kuin pystysarakkeen menetelm .

4.2 Relevantit dokumentit

Relevanttien dokumenttien korpuksessa tehtyjen hakujen tulokset esitetään kuviossa 3 ja taulukossa 3. Tälläkin relevanssitasolla rakenteinen tesauruslaajennus oli kaikilla saantitasoilla tehokkain kyselytyyppi ja ainoa peruskyselyä tehokkaampi.

Huonoin menetelmä on 10-70 % saannilla litteä tesauruslaajennus ja korkeammilla tasoilla litteä Finthes-laajennus. Ainoa peruskyselyä parempi keskiarvotarkkuus on rakenteisella tesauruslaajennuksella (ero 6,3 prosenttiyksikköä), mutta erolla ei Sparck Jonesin mukaan ole käytännössä suurta merkitystä. Tilastollisesti ero on melko merkitsevä (p<0,05 Relevanttien dokumenttien Friedmanin testin tunnusluku on

2×10^{-9} eli menetelmien väliset erot ovat erittäin merkittäviä.

Sparck Jonesin peukalotuntumalla tärkeä ero on vain litteän ja rakenteisen tesauruslaajennuksen välillä, 11,1 prosenttiyksikköä. Tilastollisesti tämä ero on varsin merkitsevä ($p<0,05$).

Tasolla kaikki relevantit havaittu tesauruksella laajentamisen erikoinen menestys vain korostuu tällä tasolla, kun suurin Friedmanin testin p-arvo löytyy litteän ja rakenteisen tesauruslaajentamisen välillä ja keskiarvotarkkuuksista litteän tesauruslaajennuksen arvo on huonoin ja rakenteisen paras.

4.3 Erittäin relevantit dokumentit

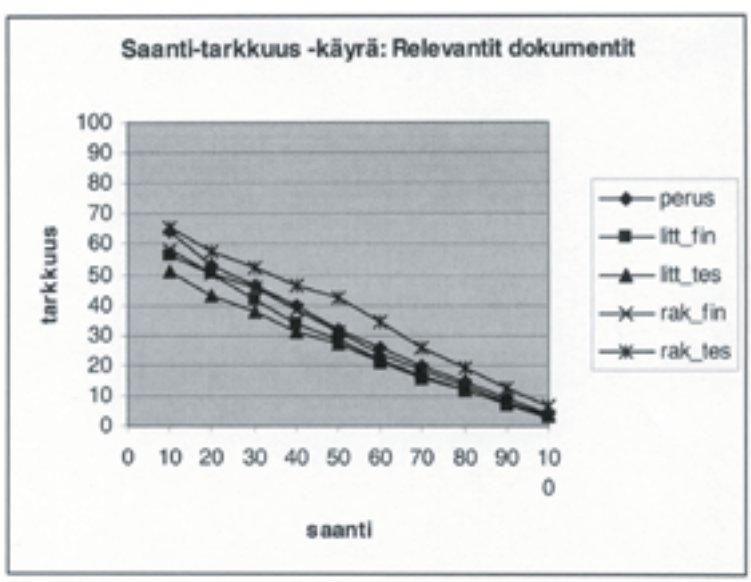
Tälläkin tasolla rakenteinen tesauruslaajennus on paras hakumenetelmä kaikilla saantitasoilla (taulukko 5, kuvio 4). Samoin kuin relevanttien dokumenttien korpuksessa litteä tesauruslaajennus on huonoin menetelmä 70 prosentin saantitasolle asti ja litteä Finthes-laajennus 80 prosentin saantitasolla. 90 ja 100 prosentin saantitasolla tämän relevanssitason heikoin menetelmä on rakenteinen Finthes-laajennus. Keskiarvotarkkuuksien häntää pitää jälleen litteä ja kärkeä rakenteinen tesauruslaajennus.

Erittäin relevanttien dokumenttien korpuksessa menetelmien välillä on entistä vähemmän eroa. Saantikantojen koko tällä tasolla on selvästi pienempi kuin muissa korpuksissa, joten yhden relevantin dokumentin löytyminen tai löytymättä jäämisellä on enemmän seurauksia

kuin suuremmissa relevanssikorpuksissa. Friedmanin testin tunnusluku oli $4,69\times10^{-5}$ eli tälläkin tasolla on silti erittäin merkitseviä eroja.

Tällä tasolla keskiarvotarkkuuksien perusteella molemmat rakenteiset kyselytyypit toimivat paremmin kuin peruskysely (rakenteinen tesauruslaajennus – peruskysely: 3,7 prosenttiyksikköä, rakenteinen Finthes-laajennus – peruskysely: 0,2 prosenttiyksikköä). Peruskyselyn ja rakenteisen tesauruslaajennuksen välinen ero on tilastollisesti melko merkitsevä ($p<0,05$, taulukko 6), mutta Sparck Jonesin mukaan ei kiinnostava.

Ero peruskyselyn ja rakenteisen Finthes-laajennuksen välillä ei ole tilastollisesti merkitsevä eikä systemaattinen eikä Sparck Jonesin peukalotuntumalla edes kiinnostava. Yli 10 prosenttiyksikön eroja tällä menetelmällä ei syntynyt yhtään ja 5-10 prosenttiyksikön eroja vain litteän ja rakenteisen tesauruslaajennuksen välille (6,3 prosenttiyksikköä). Molempien litteiden menetelmien huonommuus raken-



Kuvio 4: Erittäin relevantit dokumentit - saanti-tarkkuus-käyrä eri kyselymenetelmillä

Taulukko 5: Erittäin relevantit dokumentit - tarkkuus saantitasoittain eri kyselymenetelmillä (paras tarkkuus varjostettu, huonoin tarkkuuus alleviivattu)

saanti	perus	litt_fin	litt_tes	rak_fin	rak_tes
10	43,2	39,7	35,5	41,5	44,4
20	34,4	32,5	26,9	34,6	35,9
30	23,7	24,7	23,4	27,2	29,7
40	20,5	21,5	19,1	23,3	27,3
50	18,7	18,8	16,4	21,9	25,4
60	14,4	14,4	13,4	14,1	18,6
70	12,3	11,7	9,6	11,5	15,3
80	9,3	7,7	7,9	8,1	12,9
90	6,7	4,7	5,4	4,5	8,7
100	5,0	3,3	4,5	3,1	6,7
avg	18,8	17,9	16,2	19,0	22,5

Taulukko 6: Friedmanin testi: erittäin relevantit

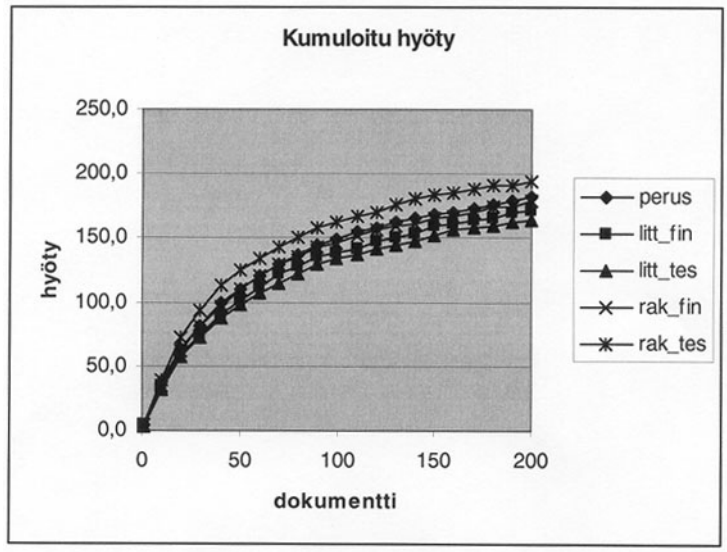
	perus	litt_fin	litt_tes	rak_fin
litt_fin	*			
litt_tes	*	-		
rak_fin	-	*	*	
rak_tes	*	***	***	*

teiseen tesauruslaajennukseen nähden on erittäin merkitsevä, mutta Sparck Jonesin peukalotuntumalla merkityksetön (litteä Finthes –laajennus, ero 4,6 prosenttiyksikköä) tai vain kiinnostusta herättävä (litteä tesauruslaajennus, ero 6,3 prosenttiyksikköä).

4.4 Kumuloitu hyöty

Kumuloitua hyötyä laskettaessa määritetään painot eri relevanssitasoille. Käytin kumuloidun hyödyn laskemisessa erittäin relevanteille dokumenteille painoa 10, relevanteille painoa 5 ja melko relevanteille painoa 1. Erittäin relevantti dokumentti oli siis kymmenen kertaa arvokkaampi kuin melko relevantti dokumentti. (ks. tarkemmin Järvelin & Kekäläinen, 2002).

Kumuloidun hyödyn perusteella lasketut tulokset eivät paljon poikenneet perinteisin menetelmin saaduista. Paras menetelmä kahdensadan dokumentin listalla on rakenteinen tesauruslaajennus kaikkien muiden, paitsi ensimmäisen dokumentin kohdalla (ks. taulukko 7). Huonoin menetelmä läpi koko listan on litteä tesauruslaajennus. Kuviosta 5 käy ilmi, että



Kuvio 5: Kumuloitu hyöty: Mentelmien erot tulosjoukoilla 1-200

rakenteisen tesauruslaajennuksen jälkeen paras menetelmä on peruskysely, mutta ihan sen kyljessä kulkee litteä Finthes-laajennus.

5. Keskustelu ja johtopäätökset

Voorheesin (1994) litteiden kyselyjen tutkimuksessa vain lyhyiden kyselyjen tulos parani merkitsevästi laajentamalla. Omat kyselyni olivat lähinnä Voorheesin lyhyiden kyselyjen pituisia. Voorheesin lyhyiden kyselyjen hakutulosta hänen käyttämänsä laajennusmenetelmä paransi merkitsevästi. Omassa kokeessani litteä laajennus kummallakaan sanastolla ei parantanut hakutulosta merkitsevästi millään relevanssitasolla. Voorheesin tutkimuksessa laajennusavaimia olivat kaikki kyselyn avaimiin suoraan liittyvät avainfasetit, siis myös ylempiä, alempia ja rinnakkaistermejä. Kekäläiselläkin (1999) paras tulos syntyi laajentamalla mahdollisimman voimakkaasti eli niin synonyymeilla, suppeammilla käsitteillä kuin rinnakkaiskäsitteillä, kun taas omassa tutkimuksessani laajennusavaimiksi valittiin vain synonyymit. Kekäläisen järjestelmä oli sama probabilistinen InQuery kuin itselläni. Voorheesin, Kekäläisen ja omien tulosteni perusteella näyttäisi siltä, että pelkät synonyymit ovat liian suppea laajennusluokka ainakin jos laajentaminen tehdään litteästi.

Yleisellä tasolla tuloksista on helppo vetää ainakin se johtopäätös, että näitä kahta laajennus menetelmää ja kahta laajennusavainlähdettä verrattaessa, jos kyselyä laajennetaan automaattisesti synonyymisanastolla, se pitää ehdottomasti tehdä rakenteisesti ja dokumenttikokoelmaa varten räätälöidyllä sanastolla. Tätä tukevat kaikkien aiheiden tarkkuuksien keskiarvot. Tässä tutkimuksessa käytetty rakenne on niin yksinkertainen, että sen automatisoiminen esimerkiksi tietokantaan liitettyä sanastoa käytettäessä ei ole vaikeaa.

Toinen yhtä itsestäänselvä tulos oli, että rakenteisesti laajennettaessa tietokantaa varten räätälöity sanasto on ehdottomasti parempi laajennusavainten lähde kuin Finthes. Tuloksiin on varmasti vaikuttanut se tutkimuksessa käytettyjen tesaurusten ero, että TUTK-tesauruksessa laajennusavaimet valittiin olettaen laajennettavien hakuavainten olevan jo perusmuodossaan, kun taas Finthes tulkitsi hakuavaimesta kaikki mahdolliset taivutusmuodot ja kantasanat. Tähän tutkimukseen valittu periaate ottaa mukaan kritiikittömästi kaikki Finthesillä ja Fintwolilla tuotetut laajennusavaimet tuo mukaan paljon semanttisesti asiaankuulumattomia

Taulukko 7: Kumuloitu hyöty: erot prosenttiyksikköinä. Taulukossa tummennettu ne ruudut, joissa vaakarivin menetelmä on parempi kuin pystyrivin mentelmä

	perus	litt_fin	litt_tes	rak_fin
litt_fin	8,8			
litt_tes	13,6	4,8		
rak_fin	2,0	6,8	11,6	
rak_tes	12,6	21,0	25,8	14,2

Taulukko 8: Kumuloitu höyty: Erot prosenttiyksikköinä. Taulukossa tummennettuna ne ruudut, joissa vaakarivin menetelmä on parempi kuin pystyrivin menetelmä

	perus	litt_fin	litt_tes	rak_fin	rak_tes
1	4,6	5,0	3,5	4,3	4,8
10	37,1	33,0	31,9	35,1	39,3
20	63,0	58,1	57,0	62,1	73,0
30	81,6	74,5	73,0	80,7	93,1
40	98,9	90,1	86,9	96,2	113,7
50	110,5	100,7	97,7	109,3	124,3
60	120,7	111,9	107,4	119,9	133,5
70	130,3	121,4	114,4	129,2	142,9
80	137,7	128,5	122,1	136,0	150,0
90	144,4	135,6	128,9	143,5	158,4
100	148,9	138,0	133,8	146,8	162,5
110	154,4	142,3	137,7	152,3	167,9
120	157,8	147,1	141,8	156,6	170,7
130	162,0	150,6	145,1	159,4	176,1
140	165,2	154,1	148,2	161,5	180,9
150	168,1	159,0	152,6	165,9	183,6
160	170,5	161,3	156,0	167,5	185,1
170	173,5	164,0	157,9	170,6	187,9
180	176,0	166,3	160,2	174,3	190,7
190	179,8	170,0	162,1	174,3	192,0
200	182,0	172,4	164,3	179,8	194,2
Keskiarvo	131,8	123,0	118,2	129,8	144,0

laajennustermejä. Jos laajennusavainten lähde tulkitsisi sanat vain perusmuodossa, huonoja laajennusavaimia ei todennäköisesti pääsisi mukaan niin paljon ja laajennuksen tulos olisi todennäköisesti parempi.

Tässä tutkimuksessa Finthesillä rakenteisesti laajentaminen oli peruskyselyä huonompi menetelmä sekä perinteisin menetelmin että kumuloidulla hyödyllä mitattaessa. Rakenteisen Finthes-laajennuksen ja peruskyselyn välinen ero ei tosin ole millään tasolla tilastollisesti merkitsevä eikä Sparck Jonesin mukaan käytännössä edes mielenkiintoinen, mutta ero peruskyselyn hyväksi on systemaattinen kaikissa muissa paitsi erittäin relevanttien dokumenttien korpuksessa. Kumuloitua hyötyä rakenteinen tesauruslaajennus tuottaa 12,6 prosenttiyksikköä enemmän kuin peruskysely, mikä on peukalosäännön mukaan jo käytännössä merkittävä ero. Mitään syytä laajentaa litteästi tai Finthesillä tämä työ ei siis löydä. Myös tesauruksella laajentamalla tuotetun rakenteisen synonyymilaajennuksen ero peruskyselyyn on niin vähäinen, että todelliseksi tiedonhakujärjestelmän parantajaksi siitä tuskin on.

Hyväksytty julkaistavaksi 1.11.2003

Lähteet:

Alaterä, A., Halttunen, K. (2002). Tiedonhaun perusteet – osa lukutaitoa. Tampereen yliopiston täydennyskoulutuskeskus ja Otavan Opisto/ Internetix. Helsinki: BTJ Kirjastopalvelu.

Applied Computing Systems Institute of Massachusetts, Inc. (ACSIOM) (1996). InQuery document retrieval system. Ohjetiedosto.

Broglio, J., Callan, J. P., Croft, W. B. (1994). INQUERY System Overview. Proceedings of the TIPSTER Text Program (Phase I). San Francisco, CA. Morgan Kauffman. 47-67. Saatavilla myös www-muodossa: <http://ciir.cs.umass.edu/pubfiles/brogliocallancrofttipI.pdf> Käytetty 12.2.2002.

Callan, J. P., Croft, W. B., Harding, S. M. (1992). The INQUERY Retrieval System. Proceedings of the 3rd International Conference on Database and Expert Systems Applications. 78-83. Saatavilla myös www-muodossa: <http://www.cs.cmu.edu/~callan/Papers/callancroftdexa92.ps.gz> Käytetty 12.2.2002.

Efthimiadis, E. (1996). Query Expansion. Annual Review of Information Science and Technology (ARIST) 31. Medford, NJ, 121-187.

Järvelin, K. (1995). Tekstitiedonhaku tietokannoista. Espoo: Suomen ATK-kustannus Oy.

Järvelin, K., Kekäläinen, J. (2002). Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems (ACM TOIS) 20(4): 422-446.

Järvelin, K., Kekäläinen, J., Niemi, T. (2001). ExpansionTool: Concept-based query expansion and construction. Information Retrieval 4(3/4), 231-255. Saatavilla myös www-muodossa Tampereen yliopiston informaatiotutkimuksen laitoksen julkaisusarjassa osoitteessa : <http://www.info.uta.fi/julkaisut>.

Kekäläinen, J. (1999). The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. Väitöskirja, informaatiotutkimuksen laitos Tampereen yliopisto. Acta Universitatis Tamperensis 678. Tampere: University of Tampere.

Kristensen, J. (1992). Vapaasanahakujen laajentaminen hakutesauruksen avulla haettaessa indeksoimattomasta tekstitietokannasta. Tampere: Tampereen Yliopisto. Kirjastotieteen ja informatiikan lisensiaattitutkielma.

Magennis, M., van Rijsbergen, C. (1997). The potential and actual effectiveness of interactive query expansion. Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. New York. NY: ACM, 324-332.

Salton, G. (1989). Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Publishing Company. Addison-Wesley Series in Computer Science.

Siegel, S. (1989). Nonparametric statistics for the behavioral sciences. New York, NY: McGraw-Hill.

Sormunen, E. (1993). Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanomalehtiaineistoa sisältävässä tekstikannassa. Tampere: Tampereen yliopisto 1993. Kirjastotieteen ja informatiikan lisensiaattitutkielma.

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

Voorhees, E. (1994). Query Expansion Using Lexical-Semantic Relations. Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM, 61-69.

Verkkolähteet

Lingsoft. Finthes-ohjelman demo. Käytetty 20.8.2002. IRL: <http://www.lingsoft.fi/cgi-bin/finthes>

Lingsoft. Fintwol-ohjelman demo. Käytetty 3.10.2002. IRL: <http://www.lingsoft.fi/cgi-bin/fintwol>

Muut lähteet

Ronkainen, O-V. (2002). Sähköpostiviesti Eija Airiolle 9.10.2002. Aihe: Finthes.

Liite 1. Hakuaiheet:

		kuvaus
1	**summit**	George Bushin ja Mihail Gorbatsovin tapaaminen Helsingissä syyskuussa 1990. Neuvotteluissa käsitellyt asiat sekä tehdyt päätökset ja sopimukset.
2	**velka**	Etelä-Amerikan velkakriisi. Miten velkaantumisongelma on kehittynyt? Miten ongelmaa on pyritty ratkaisemaan?
3	**polku**	Metsäteollisuuden polkumyyntisyytteet USA: ssa. Kiinnostavaa suomalaisten paperinviejien kohtalo. Polkumyyntisyytösten sisältö, oikeudenkäynnin tulokset.
4	**jykul**	Jyväskylän kaupungin ja maalaiskunnan kuntaliitoshanke. Halutaan kartoittaa liitoshankkeen kannattajien ja vastustajien mielipiteitä ja perusteluja. Arviot liitoksen taloudellisista vaikutuksista (mm. porkkanaraha).
5	**varso**	Varsovan liiton lakkauttaminen. Mitä tahansa muutosprosessista, eri jäsenmaiden suhtautumisesta, päätöksistä jne.
6	**liett**	Neuvostoliiton Liettuaan kohdistama taloussaarto keväällä 1990. Mitä toimia taloussaartoon liittyi ja miten se näkyi Liettuassa? Saarron lopettamiseen johtaneet tapahtumat.
7	**iraki**	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?
8	**opec**	OPEC:n öljyn hintaa ja tuotantomääriä koskevat päätökset.
9	**bukar**	Presidentti Iliescun hallituksen avuksi kutsumien kaivosmiesten väkivaltaisuudet oppositiota vastaan Bukarestissa. Taustatietoja tapahtumista, uhreista ja jälkiselvittelyistä.
10	**untag**	Namibian itsenäistymiseen liittynyt YK:n rauhanturvaoperaatio. Tietoja operaation valmistelusta, siihen liittyneistä tapahtumista sekä UNTAG-joukkojen ja sen suomalaispataljoonan toiminnasta.
11	**eyval**	EY:n parlamentin asema yhteisön päätöksenteossa. Halutaan selvittää EY:n parlamentin asema suhteessa komissioon ym. toimielimiin. Mitä muutoksia nykyiseen on haluttu ja ketkä ovat halunneet? Miten demokraattinen kontrolli toimii EY: ssä?
12	**bildt**	Carl Bildt ja pohjoismainen yhteistyö. Bildtin pohjoismaista yhteistyötä koskevat lausunnot. Mitä erityistä Bildt on sanonut Ruotsin ja Suomen yhteistyöstä?
13	**jugos**	Jugoslavian presidenttineuvoston toimintaa koskevat uutiset. Erityisesti tiedot istunnoista ja niissä tehdyistä päätöksistä.
14	**saksa**	Länsi- ja Itä-Saksan sekä miehittäjävaltioiden (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?

15	**valmet**	Valmetin traktori- ja kuljetusvälinetuotannon kannattavuus. Kuljetusvälinetoimialaan lasketaan kuuluvaksi metsä- ja siirtokoneet sekä kiskokalusto (mm. Transtech). Osakkuudet henkilö- ja kuorma-autoteollisuudessa jätetään tarkastelun ulkopuolelle.
16	**tampel**	Tampellan irtisanomiset. Tavoitteena koota tietoja Tampella-konserniin kuuluvien yhtiöiden suorittamista irtisanomisista. Tietoja lomautuksista ja lyhennetyistä työviikoista ei tarvita.
17	**matka**	Keran ja KTM:n investoinnit matkailuun. Tietoja matkailualan yrityksille myönnetyistä avustuksista ja lainoista (=tässä investointi). Erityisen arvokkaita yhteenvedot.
18	**neste**	Neste Oy:n maakaasutoiminta. Halutaan yleiskuva Nesteen maakaasutoiminnoista. Mitä Neste on puuhaillut maakaasun hankinnan (kentät ja tuontisopimukset), jakelun (verkoston rakentaminen) ja markkinoinnin alueilla.
19	**yjate**	Ydinvoimalaitosten tuottamien radioaktiivisten jätteiden käsittely ja varastointi. Esimerkkejä ongelmista, riskeistä ja sattuneista ydinjätevahingoista.
20	**aids**	AIDSin levinneisyys EY-maissa. Miten vakava AIDS-tilanne on näissä maissa? Tietoja esiintymämääristä ja kampanjoista ym. taudin leviämistä ehkäisevistä toimista.
21	**elint**	Elintarvikkeiden tuontirajoitukset ja –säännöstely eri maissa. Rajasuojan ja sen vähentämisen vaikutus elintarviketeollisuuteen erityisesti Suomessa. Selvityksiä, arvioita, mielipiteitä ym. taustatietoa.
22	**asunt**	Asuntotuotannon suhdanteet ja suhdannevaihtelut Suomessa; erityisesti tilasto- ja ennustetietoja, arvioita.
23	**paast**	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.
24	**japan**	Japanin autoteollisuuden investoinnit Eurooppaan ja tuotannollinen yhteistyö eurooppalaisten autonvalmistajien kanssa. Mihin maihin japanilaisia autotehtaita on suunniteltu, perustettu ja laajennettu? Tuotantomärät ja –trendit.
25	**sellu**	Metsäteollisuuden ympäristöinvestoinnit. Rajoitutaan vesiensuojeluun liittyviin investointeihin kemiallisessa metsäteollisuudessa. Sekä varsinaiset puhdistamoinvestoinnit että ymp äristöystävällisempien prosessien käyttöönotto.
26	**aukio**	Kaupan aukioloajat. Halutaan selvittää vähittäiskauppojen aukioloaikojen vapauttamista koskevaa keskustelua. Erityisesti kartoitetaan kaupan järjestöjen ja ammattijärjestöjen kannanottoja ja toimia.
27	**kierr**	Pakkaukset ympäristönsuojelukys ymyksenä. Erityisesti kiinnostavat kulutustavarapakkausten kierrätysjärjestelmät, niiden kehittämiskokeilut, kierrätykseen liittyvä lainsäädäntö eri maissa.
28	**eyaho**	Esko Aho ja Suomen EY-jäsenhakemus. Ahon Suomen EY-jäsenyyden hakemiseen liittyvät mielipiteet, kannanotot ja toimet. Muiden arviot Eskon toimista ja puheista.
29	**ydinv**	Kauko Juhantalon ydinvoimapuheet ja –teot. Juhantalon perustelut 5. ydinvoimalan puolesta. Miten Juhantalo vei ydinvoimalaratkaisua eteenpäin?
30	**vihr**	Vihreiden tekemät aloitteet, välikysymykset, ehdotukset, puheenvuorot ja äänestyskäyttäytyminen Suomen eduskunassa. Tarkastelussa sekä ryhmä että yksittäiset kansanedustajat.