KYSELYN AUTOMAATTINEN LAAJENTAMINEN SYNONYYMEILLA

Sari Kaitaniemi 65133
Tampereen yliopisto
informaatiotutkimuksen laitos
ohjaaja K Järvelin
20.11.2002

TIIVISTELMÄ

Tampereen yliopisto

Informaatiotutkimuksen laitos

KAITANIEMI, SARI: Kyselyn automaattinen laajentaminen synonyymeilla

Pro gradu -tutkielma, 74 sivua ja 18 liitesivua

Tiedonhaku

Marraskuu 2002

Tämän tutkielman aihe on suomenkielisen kyselyn automaattinen laajentaminen synonyymeilla probabilistisessa hakujärjestelmässä. Tutkielman tietokanta on suomenkielinen TUTK, joka käsittää 54 000 artikkelia suomalaisista sanomalehdistä. Hakujärjestelmä on probabilistinen InQuery. Synonyymien lähteenä käytetään kahta sanastoa: yleistä, kaupallista Finthes-synonyymisanastoa ja TUTKia varten räätälöityä tesaurusta. Kyselyt laajennetaan sekä rakenteettomasti että yksinkertaisella rakenteella, jossa synonyymifasetit on yhdistetty syn-operaattorilla. Ennen laajennuskokeita verrataan kahta sanaliiton käsittelymenetelmää: sanaliiton kaikki osat yhdistetään syn-operaattorilla tai sanaliiton osien synonyymit yhdistetään syn-operaattorilla ja fasetit toisiinsa uwn-läheisyysoperaattorilla. Sanaliittojen käsittelyssä syn+uwn-menetelmä osoittautui hieman paremmaksi kuin syn-menetelmä.

Tietokannan relevanssiarviot on tehty neliportaisesti: ei relevantti, vähän relevantti, melko relevantti ja erittäin relevantti. Kokeet tehdään kolmella tasolla. Ensimmäisessä korpuksessa ovat mukana kaikki relevantit dokumentit. Toisessa korpuksessa ovat mukana melko ja erittäin relevantit dokumentit. Pienimmässä korpuksessa on vain erittäin relevantit dokumentit.

Tuloksia verrataan saanti—tarkkuus -käyrillä ja -taulukoilla, tarkkuuksien keskiarvoilla sekä kumuloidulla hyödyllä ja alennetulla kumuloidulla hyödyllä. Tulosten tilastollista merkitsevyyttä mitataan Friedmanin testillä ja Karen Sparck Jonesin prosenttiyksikkömääräisiin eroihin perustuvalla peukalosäännöllä.

Ainoa menetelmä, joka oli kahdessa relevanssikorpuksessa ja kumuloiduilla menetelmillä parempi kuin laajentamaton peruskysely, oli rakenteinen tesauruksen synonyymeilla laajennettu kysely. Molemmat litteät menetelmät olivat kaikissa kolmessa korpuksessa huonompia kuin peruskysely, tilastollisesti joko melko tai varsin merkitsevästi. Kumuloidun hyödyn menetelmät vahvistavat peruskyselyn paremmuutta. Rakenteisen Finthes-laajennuksen ja peruskyselyn välinen ero ei ole tilastollisesti merkitsevä missään korpuksessa, mutta kahden laajemman korpuksen keskiarvotarkkuus ja molemmat kumuloidut menetelmät osoittavat peruskyselyn olevan rakenteista Finthes-laajennusta parempi menetelmä.

Mitään syytä laajentaa litteästi tai Finthes-sanastolla tämä työ ei löydä. Ainoa hyödyllinen synonyymilaajennus on rakenteinen, tekstikokoelmaa varten räätälöidyllä sanastolla tehty laajennus. Muiden tutkimusten tuloksiin yhdistettynä tämä tutkielma osoittaa, että pelkkä synonyymilaajennus ei liene riittävän tehokas laajennusmenetelmä, vaan laajennusavainpohjan tulisi olla kattavampi.

KYSELYN AUTOMAATTINEN LAAJENTAMINEN SYNONYYMEILLA 1

1. Johdanto

Tämä työ kuuluu tiedonhaun tutkimukseen. Tiedon hankkiminen on yksi ihmisen elämän perustoiminnoista. Arkisesti haetaan puhelinnumeroita, hintoja – jopa juoruja. Koiraa hankkiva ihminen haluaa tietää mahdollisimman paljon itseään kiinnostavista roduista. Vastasairastunut ihminen tarvitsee tietoa sairaudestaan ja sen hoitovaihtoehdoista. Tietoyhteiskunnassa tiedonhankinta kuuluu automaattisesti moneen toimenkuvaan. Tietoa tarvitseva ihminen kysyy aiheesta yleensä ensin läheisiltään, tuttaviltaan tai kollegoiltaan. Vasta jos apu ei löydy lähipiiristä, tietoa haetaan virallisemmista lähteistä, kuten kirjastosta, tietopalvelusta ja nykyään Internetistä tai kaupallisista tietokannoista (Wilson 1977, 45).

Tiedonhaku tutkimusalana käsittää informaatiota sisältävien kohteiden esittämisen, varastoimisen, järjestämisen ja niiden saatavuuden varmistamisen. Informaatiota sisältävän kohteen ulkomuodolle on vaikea asettaa rajoituksia. (Salton & McGill 1983, 1.) Arvokasta informaatiota voi löytyä paitsi kirjoista, artikkeleista ja arkistoista – myös webbisivulta, museoesineestä tai luonnonsuojelualueelta.

Luonnollinen kieli rikastuttaa ihmisten välistä kommunikointia, mutta rikkaus 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 asioille voi olla useita eri nimityksiä, synonyymejä. Lisäksi asiat voidaan ilmaista joko yksityiskohtaisesti tai yleisellä tasolla. Tässä työssä tarkoitukseni on tutkia tiedonhakua suomenkielisestä artikkelitietokannasta. Tutkimusongelma on: miten suomenkielisen kyselyn automaattinen laajentaminen synonyymeilla vaikuttaa hakutulokseen probabilistisessa hakujärjestelmässä. Tutkimus on empiirinen laboratoriotutkimus, jonka tarkoitus on evaluoida kyselynlaajentamismenetelmiä.

Tietoa hakee yhä useammin tiedon tarvitsija itse, kun kaupalliset ja Internetin tietojärjestelmät tuovat tiedon tarvitsijoiden ulottuville. Ammattitaitoisten välittäjien käyttäminen tiedonhaussa vähenee. Kokematon tiedonhakija tekee usein lyhyitä kyselyitä. Hän ei välttämättä hallitse kyselynmuodostuksen tekniikoita eikä analysoi aihettaan niin huolellisesti, että osaisi muodostaa tyhjentävän kyselyn. Äkkiseltään vuorovaikutteinen kyselyn laajentaminen voisi tuntua hyvältä ratkaisulta; annetaan hakijan itse valita sanastosta laajennusavaimia kyselyynsä. Magennis &

Rijsbergenin (1997, 74-81) mukaan kokemattomat käyttäjät eivät yleensä osaa valita hyödyllisiä laajennusavaimia ja hakutulos useimmin ei ainakaan parane vuorovaikutteisella laajentamisella. Satunnaista tiedonhakijaa auttaisi todennäköisesti parhaiten järjestelmä, joka osaa itse laajentaa kyselyä tarkoituksenmukaisesti.

Laajennusavaimet voidaan ottaa monista lähteistä: omasta päästä, tekstikokoelmasta itsestään, intellektuaalisesti tiettyä tietokantaa ja tiedonhakua varten rakennetusta sanastosta tai yleisestä, kaupallisesta sanastosta. Sanaston rakentaminen on kallista ja aikaa vievää työtä, joten helpoin ratkaisu olisi käyttää jotakin saatavilla olevaa valmista synonyymisanastoa. Tässä työssä verrataan kahta eri lähdettä: kaupallista, yleistä Finthes-synonyymisanastoa sekä tesaurusta, joka on rakennettu tutkimuksen tietokantaa ja siitä tapahtuvaa tiedonhakua varten. Kyselyjen kaikki hakuavaimet laajennetaan. Koska Finthes antaa synonyymejä myös taivutetuissa muodoissa, Finthesin antamat avaimet perusmuotoistetaan Fintwol-ohjelmalla. Laajennoksista tehdään sekä rakenteeton (litteä) että rakenteinen kysely ja niiden tuloksellisuutta verrataan.

Lisäksi verrataan kahta tapaa käsitellä sanaliittoja InQuery-hakujärjestelmässä. Ensimmäisessä menetelmässä sanaliiton kaikki osat synonyymeineen yhdistetään yhtenä ryppäänä syn-operaattorilla. Toisessa menetelmässä sanaliiton osat yhdistetään läheisyysoperaattorilla uwn ja sen sisällä synonyymit yhdistetään syn-operaattorilla.

1.1 Työn rakenne

Ensimmäisessä luvussa käsittelen kyselynlaajentamisen keskeisiä käsitteitä sekä yleisesti että tämän työn kannalta. Toinen luku käsittelee luonnollisesta kielestä tiedonhakuun johtuvia ongelmia. Luvussa kolme tutustutaan lyhyesti erilaisiin hakujärjestelmiin ja niiden pohjalla oleviin täsmäytysmenetelmiin. Luku neljä luo lyhyen katsauksen aihealueen tähänastiseen tutkimukseen. Luvussa viisi esittelen tekemäni tutkimuksen kokoelman, hakujärjestelmän, synonyymilähteet, kyselyt ja menetelmät. Luvussa kuusi selvitän kokeideni tuloksia. Luku seitsemän sisältää keskustelua tuloksista ja johtopäätöksiä aiemman ja tämän tutkimuksen pohjalta.

1.2 Käsitteet

Dokumentti

Dokumentti on tietovälineen ja siihen tallennetun tiedon muodostama asiasisällöltään rajattu

kokonaisuus (Tietohuollon sanasto 1993, 13). Tiedonhakujärjestelmässä dokumentti on yleensä tallennettu digitaaliseen muotoon. Se voi olla paitsi tekstiä, myös kuva, ääntä tai videokuvaa. Käyttämissäni lähteissä dokumentista on käytetty myös nimityksiä information object ja information item. Nämä kaikki olen kääntänyt sanalla dokumentti.

Indeksointi

Indeksointi on sopivien kuvailutermien eli indeksiavainten liittämistä kokoelman dokumentteihin. Jos sisällönkuvailu tehdään automaattisesti, puhutaan automaattisesta indeksoinnista. Ihmisen tekemä sisällönkuvailu on intellektuaalista indeksointia. Automaattisessa indeksoinnissa dokumentin on oltava digitaalisessa muodossa. Dokumenteista luodaan esitysmuoto, jossa on dokumenttia kuvailevaa, objektiivista tietoa (esimerkiksi tekijä, julkaisuvuosi jne) ja sisältöä kuvailevia indeksiavaimia. (Salton & McGill 1983, 52-55.)

Indeksoinnilla on kolmitahoinen tavoite: tiedonhakijaa kiinnostavien dokumenttien paikantaminen, dokumenttien ja aihealueiden suhteuttaminen toisiinsa ja dokumentin relevanssin määrittäminen tietyn kyselyn suhteen. Indeksiavaimet voidaan valita joko kontrolloidusta tai kontrolloimattomasta sanastosta. Kontrolloitu sanasto on tietojärjestelmään liittyvä valmis sanasto, esimerkiksi tesaurus. Kontrolloimaton sanasto saa avaimensa dokumenttien teksteistä ja kyselyistä. Dokumentin sisältö voidaan kuvailla joko yksittäisin avaimin tai käyttämällä avaimia kontekstissaan. (Salton & McGill 1983, 54.)

Esimerkiksi voidaan ajatella minimaalista neljän sanan indeksointikieltä. Kun kielen indeksiavain sopii kuvaamaan dokumentin sisältöä, dokumenttivektorissa avainelementin arvo on 1. Jos indeksiavain ei sovi kuvaamaan dokumenttia, vektorissa avaimen kohdalla on 0. Jos dokumenttia voidaan kuvata kaikilla näillä indeksiavaimilla, dokumenttia kuvaa vektori

<1 1 1 1>.

Jos dokumenttia kuvaavat vain avaimet 1 ja 4, vektori on tämän näköinen:

<1 0 0 1>.

(Salton & McGill 1983, 12.)

Hakuavain

Tässä työssä tarkoitan termillä hakuavain niitä sanoja ja sanojen osia, joilla tiedonhakija kuvaa tiedontarvettaan ja joista (mahdollisesti operaattorein yhdistämällä) syntyy kysely. Muissa töissä hakuavaimesta käytetään myös nimityksiä hakusana ja hakutermi.

Kysely

Relevantteja dokumentteja löytääkseen tiedonhakijan tulee pystyä antamaan kaikki ne – ja vain ne – hakuavaimet, joilla tiedon tarpeen tyydyttävät dokumentit on kuvailtu. Tiedon tarpeen selkiytyessä tarvitsija muotoilee tiedontarpeensa hakupyynnöksi, jonka joko hän itse tai tiedon välittäjä muokkaa kyselyksi. Tiedonhaun tuloksen kannalta parhaassa kyselyssä esiintyvät samat avaimet, kuin relevanttien dokumenttien kuvailussa on käytetty (Kekäläinen, 1999). Siihen, kuinka tiedon tarvitsija ilmaisee tiedontarpeensa, vaikuttaa hänen viitekehyksensä, taustansa ja aikaisempi tietämyksensä (Swanson, 1988). Tiedon tuottajan ja/tai indeksoijan viitekehys saattaa olla erilainen, joten käytetyt indeksointi- ja hakuavaimet eivät välttämättä kohtaa.

Kyselyt voivat olla joko rakenteisia tai rakenteettomia eli litteitä. Rakenteisissa kyselyissä hakuavaimien väliset suhteet ilmaistaan operaattoreilla, joita ovat esimerkiksi Boolen logiikan perusoperaattorit and, or ja not. Samaa käsitettä edustavat avaimet yhdistetään disjunktiolla eli or-operaattorilla, eri käsitteitä kuvaavat avaimet tai avainfasetit yhdistetään konjunktiolla eli and-operaattorilla. Negaatio eli not-operaattori sulkee pois kaikki dokumentit, joissa esiintyy negaatiolla merkitty hakuavain. Sulut kertovat operaattorin vaikutusalueen. Esimerkiksi jos hakija haluaa tietoa koiranpennun hampaiden vaihtumisesta, hän voi muotoilla kyselyn

> (koira **or** koiranpentu) **and** (hammas **or** maitohammas) **and** (vaihtuminen **or** irtoaminen).

Sulut vaikuttavat siten, että ensin järjestelmä käsittelee sulkujen sisällä olevat lausekkeet ja sen jälkeen koko hakulauseen. Ensin siis haetaan kolme dokumenttiryhmää, jotka koostuvat dokumenteistä, joiden indeksiavaimena esiintyy 1) koira tai koiranpentu, 2) hammas tai maitohammas ja 3) vaihtuminen tai irtoaminen. Tämän jälkeen and-operaattorin vaikutuksesta näistä kolmesta joukosta tehdään leikkaus eli lopulliseen tulosjoukkoon pääsevät dokumentit, jotka täyttävät kaikkien kolmen osalausekkeen ehdot.

Jos halutaan rajata pois tapaturmaisesti irronneita hampaita käsittelevät dokumentit, kysely voidaan muotoilla:

> ((koira **or** koiranpentu) **and** (hammas **or** maitohammas) **and** (vaihtuminen **or** irtoaminen)) **not** (onnettomuus **or** tapaturma).

Negaatiota kannattaa käyttää varovasti, sillä se voi pudottaa tulosjoukosta myös hyödyllisiä dokumentteja. Samassa dokumentissa voidaan käsitellä sekä ei-toivottua aihetta että juuri sitä aihetta, mistä tietoa halutaan. Esimerkiksi tapaturmainen hampaiden menetys saattaa olla yksi aihe dokumentissa, jossa kerrotaan myös pennun maitohampaiden vaihtumisesta. (Ks. esim. Järvelin 1995, 142-145.)

Relevanssi

Relevanssi on tiedonhaun keskeisimpiä käsitteitä. Se on ihmiselle intuitiivisesti helppotajuinen mutta tarkemmin määriteltäessä monitahoinen ja häilyvä käsite. Tiedonhaun tutkimuksessa relevanssia pidetään järjestelmän dokumenttien ominaisuutena suhteessa kyselyyn. Tiedonhankinnan tutkimuksessa relevanssi liittyy pikemminkin käyttäjän kognitiivisiin prosesseihin ja kontekstin aiheuttamiin tietämyksen muutoksiin ja tiedontarpeisiin (Cosijn & Ingwersen, 2000). Relevanssin käsite liittyy siis vahvasti tiedon tarvitsijaan, ihmiseen. Inhimillisen tekijän vuoksi suure on vaikeasti mitattavissa.

Saracevic (1996) esitti uudenlaisen relevanssin tarkastelukulman: tiedonhaun vuorovaikutusmallin. Malli perustuu relevanssin intuitiiviseen ymmärtämiseen, filosofiaan ja kommunikaatioteoriaan. Hän on löytänyt relevanssille viisi **ominaisuutta**: suhde (relation), aikomus (intention), konteksti (context), päätelmä (inference) ja vuorovaikutus (interaction).

> **Suhde:** Relevanssilla arvioidaan esimerkiksi tiedonlähteen ja -tarpeen välistä suhdetta.
>
> **Aikomukset** vaikuttavat relevanssiin ja sen pohjalla olevaan suhteeseen. Aikomukset voivat olla tavoitteita, rooleja tai odotuksia. Myös motivaatio vaikuttaa relevanssiin.
>
> **Konteksti** määrittää aina tiedonhakua. Aikomus kohdistuu kontekstiin, tiedontarpeen aiheuttaneeseen tilanteeseen ja sen ympäristöön. Relevanssia ei voi arvioida kontekstin ulkopuolella.

Päätelmä: Suhteen hyvyydestä tehtävä arvio kuuluu relevanssiin. Arvioidaan tietyssä kontekstissa aikomuksen perusteella etsittävää informaatiota.

Vuorovaikutus muokkaa relevanssiarviota. Päätelmä syntyy dynaamisessa, vuorovaikutteisessa prosessissa, kun muiden ominaisuuksien tulkinnat muuttuvat etsijän tietotilan muuttuessa.

Toisin sanoen kognitiivisena ilmiönä relevanssiin kuuluu vuorovaikutteinen, dynaaminen päätelmä suhteesta, joka on ilmennyt tietyssä kontekstissa syntyneen aikomuksen pohjalta. Relevanssi voidaan nähdä kriteerinä, joka heijastaa kontekstissa tapahtuvaa ihmisten tai ihmisen ja dokumentin välisen tiedonvaihdon tehokkuutta kommunikatiivisessa suhteessa. (Saracevic, 1996.)

Saracevicin (1996) mukaan relevanssin **ilmenemismuodot** voidaan jakaa seuraaviin pääluokkiin niiden ilmaiseman suhteen perusteella:

Järjestelmä- tai algoritmirelevanssi kuvaa kyselyn ja dokumentin välistä suhdetta. Kyselyä vastaava dokumentti joko löytyy tai ei löydy tietokannasta järjestelmän käyttämällä menetelmällä tai algoritmilla.

Aiherelevanssi mittaa kyselyn ilmaiseman ja tekstin kattaman aiheen välistä suhdetta. Oletetaan, että sekä kyselyn että tekstin aihe voidaan nimetä.

Kognitiivinen relevanssi eli asiaankuuluvuus kuvaa tiedontarvitsijan tietämyksen ja kognitiivisen tilan ja löydettyjen tekstien välistä suhdetta. Kognitiivinen vastaavuus, informatiivisuus, uutuus, tiedon laatu jne. ovat määreitä, joilla kognitiivista relevanssia kuvataan.

Tilannerelevanssi eli käytettävyys kuvaa tilanteen, tehtävän tai ongelman ja löydettyjen tekstien välistä suhdetta. Hyödyllisyys päätöksenteossa ja epävarmuuden väheneminen kertovat tilannerelevanssista.

Motivaatio- eli affektiivinen relevanssi syntyy tiedontarvitsijan aikomusten, tavoitteiden ja motivaatioiden sekä löydettyjen tekstien välisestä suhteesta. Tyydytys, menestys ja saavutus kuvaavat motivaatiorelevanssia.

Ilmenemismuodot esiintyvät yhdessä ja vaikuttavat toisiinsa jatkuvasti. Esimerkiksi aiherelevanssiin viitataan useimmin löydettyjen dokumenttien eli järjestelmärelevanssin perusteella. Samoin kognitiivinen ja tilannerelevanssi johtuvat muista tyypeistä. Muiden relevanssien toteaminen riippuu motivaatiorelevanssista. Tiedonhakijan alkuperäinen motivaatio vaikuttaa siis siihen, kuinka muita relevanssin ilmenemismuotoja koetaan. (Saracevic, 1996.)

Cosijn & Ingwersen (2000) pohtivat Saracevicin (1996) määrittämiä relevanssin ominaisuuksia ilmenemismuotojen kannalta. Cosijn ja Ingwersen toteavat, että relevanssin ominaisuudet toimivat eri tavoin eri ilmenemismuodoissa. Lisäksi heidän mukaansa motivaatio- tai affektiivinen relevanssi on oikeastaan lineaarinen asteikko, joka kulkee objektiivisesta subjektiiviseen relevanssiin. Saracevicin määrittelemä ilmenemismuoto motivaatiorelevanssi on Cosijn & Ingwersenin mukaan sama asia kuin aikomus-ominaisuus, ja sen voisi heidän mielestään korvata sosiokognitiivisella relevanssilla. Affektiivinen relevanssi vaikuttaa heidän mielestään relevanssin kaikkiin subjektiivisiin ilmenemismuotoihin (aihe-, kognitiivinen, tilanne- ja sosiokognitiivinen relevanssi). Heidän käsityksensä relevanssin ominaisuuksien ja ilmenemismuotojen suhteesta on esitetty taulukossa 1. Affektiivinen relevanssi on ulottuvuus samalla tavalla kuin aika. Ajan vaikutus relevanssiarvioihin kasvaa vuorovaikutuksen kuluessa. Sosiokognitiivinen relevanssi on relevanssin subjektiivinen ilmenemismuoto, jossa relevanssia arvioi yksilö vuorovaikutuksessa yhteisön muiden toimijoiden kanssa. (Cosijn & Ingwersen, 2000.)

Cosijn & Ingwersenin mukaan tiedonhaun tutkimuksessa algoritmi- ja aiherelevanssia on käytetty lähinnä osittaistäsmäyttävissä järjestelmissä, kun taas aiherelevanssia ja kognitiivista relevanssia on suosittu Boolen järjestelmiin perustuvissa vuorovaikutuksen tutkimuksissa. (Cosijn & Ingwersen, 2000.) Tässä työssä relevanssi perustuu aiheenmukaisuuteen.

Taulukko 1: Relevanssin ominaisuudet ja ilmenemismuodot (Cosijn & Ingwersen 2000)

Relevanssin ominaisuudet	**Relevanssin ilmenemismuodot**				
		⇔ **Affektiivinen relevanssi** ⇔			
	Algoritminen	**Aiherelevanssi**	**Kognitiivinen**	**Tilannerelevanssi**	**Sosiokognitiivinen**
Suhde	Kysely => dokumentit	Kyselyn => dokumentin aihe	Tietämys/ tiedontarve => dokumentit	Tilanne, tehtävä, ongelma => dokumentit	Sosiokulttuurisessa kontekstissa koettu tilanne, tehtävä tai ongelma => dokumentit
Aikomus	a) Järjestelmä-riippuvainen b) Järjestelmän tarkoitus	a) Käyttäjän / arvioijan odotukset b) Kyselyn tarkoitus	Erittäin henkilökohtainen ja subjektiivinen, liittyy tiedon-tarpeeseen, aikomuksiin ja motivaatioon	Erittäin henkilö-kohtainen ja subjektiivinen, jopa tunteellinen. Liittyy tavoitteisiin, aikomuksiin ja motivaatioon	Henkilökohtainen, subjektiivinen / organisaation strategia. Liittyy tiedontarvitsijan kokemukseen, perinteisiin, tieteelliseen paradigmaan
Konteksti	Hakukoneen säätäminen	Kaikki subjektiivisen relevanssin lajit ovat kontekstiriippuvaisia			
Päätelmä	Painotus- ja relevanssilajit-telufunktiot	Aboutnessin tulkinta ja aihekysymys semanttisella tasolla	Kognitiivisen /pragmaattisen tulkinnan ja valinnan subjektiivinen ja yksilöllinen prosessi	Tiedontarvitsijan kyky hyödyntää dokumentteja hänelle merkityksellisellä tavalla	Tiedontarvitsijan (tai –ryhmän) kyky hyödyntää dokumentteja ympäristölle merkityksellisellä tavalla
Vuorovaikutus	Automaattinen relevanssipalaute tai kyselyn muokkaaminen	Relevanssi-arviot ovat sisällöstä riippuvaisia	Relevanssiarviot ovat sisältö-, piirre- ja esityksestä riippuvaisia	Sisältää vuorovaikutuksen ympäristön kanssa	Sisältää vuorovaikutuksen ympäristön sisällä
		Lisääntyvä aikariippuvuus =>			

Tiedonhaku

Ingwersenin (1996, tässä Cosijn & Ingwersen, 2000) mukaan tiedonhakuun liittyy kolme keskeistä elementtiä: järjestelmä, käyttäjät ja ympäristö.

> **Järjestelmä** sisältää tiedostoiksi järjestettyjä dokumentteja, joita järjestelmän algoritmi täsmäyttää kyselyihin.
>
> **Käyttäjällä** on ongelmasta tai työtehtävästä johtuva tiedontarve.
>
> **Sosio-organisationaalinen ympäristö** luo kontekstin, joka vaikuttaa käyttäjän toimintaan. (Cosijn & Ingwersen, 2000.)

Tiedonhaun tutkimuksessa tutkitaan hakua järjestelmistä, joissa informaatio on dokumenteissa. Tiedonhakujärjestelmä on englanniksi Information Retrieval System, (IR System). Muita informaatiota sisältäviä tietojärjestelmiä ovat esimerkiksi tiedonhallintajärjestelmä (Data Base Management System, DBMS), johdon tietojärjestelmä (Management Information System, MIS), päätöstukijärjestelmä (Decision Support System, DSS) ja kysymys-vastaus järjestelmä (Question-Answering System, QA). (Salton & McGill 1983, 7.)

> **Tiedonhakujärjestelmä** esittää, varastoi ja hakee dokumentteja tai niiden esitysmuotoja. Syötteenä annetaan kysely rakenteisena, litteänä tai luonnollisella kielellä. Kyselyn tuloksena saadaan kyselyä vastaava viitejoukko tai nykyisin yleensä dokumenttijoukko. (Salton & McGill 1983, 7-8.)
>
> **Tiedonhallintajärjestelmässä** varastoidaan, hallitaan ja haetaan yksittäisiä faktoja, jotka on tallennettu taulukon muotoon. Tietueet on jaettu kenttiin, joista kukin sisältää tietyntyyppisen tietoalkion. Esimerkiksi yrityksen henkilöstötietokannan taulukon sarakkeisiin on tallennettu henkilönumero, työntekijän nimi ja palkka. Haulla voidaan saada esimerkiksi listaus työntekijöiden sukunimistä tai palkoista, tai yhden työntekijän kaikista tiedoista. (Salton & McGill 1983, 8.)
>
> **Johdon tietojärjestelmä** on johtajien tiedon tarpeita varten kehitelty tietokanta. Johtajat saattavat tarvita päätöksenteossaan tavallisista tietokannoista poikkeavalla tavalla muokattua informaatiota. (Salton & McGill 1983, 8-9.)

Päätöstukijärjestelmillä on mahdollista yhdistellä elementtejä tiedonhakujärjestelmistä, tietokannoista ja tietokonegrafiikkajärjestelmistä päätöksenteon apuvälineeksi. (Salton & McGill 1983, 9.)

Kysymys-vastausjärjestelmä tuottaa faktatietoa luonnollisella kielellä. Tietokanta sisältää faktoja tietyltä alalta ja yleistietoa henkilöiden välisestä keskustelusta. Kysymyksen voi esittää luonnollisella kielellä. Järjestelmä analysoi kyselyn, vertaa kyselyä tallennettuun tietämykseen ja muotoilee vastauksen ilmeisen relevanteista faktoista. (Salton & McGill 1983, 9.)

Tiedonhakujärjestelmä

Tiedonhakujärjestelmä koostuu dokumenteista, kyselyistä ja menetelmästä, jolla päätetään, vastaako dokumentti kyselyä. Dokumentit voivat olla esimerkiksi ääntä, kuvaa tai videofilmiä tekstin tai dokumenttiviitteiden lisäksi. Järjestelmä muokkaa dokumentit ja kyselyt järjestelmälle käyttökelpoiseen muotoon. Dokumenttien ja kyselyjen esitysmuotoja vertaamalla täsmäytysmenetelmä tunnistaa ne dokumentit, jotka ovat relevantteja kyselyn suhteen. (Salton & McGill 1983, 10-12.)

Kokoelman dokumentit on järjestetty tiedostoiksi. Yksinkertaisin tiedostorakenne on lineaarinen jono. Se on järjestämätön dokumenttikokoelma. Tiedonhaku on dokumenttien selaamista yksitellen. Lineaarinen jono tarvitsee vain vähän tilaa ja sen ylläpito on helppoa, kun uudet tietueet vain lisätään jonon jatkoksi. Järjestetyssä peräkkäistiedostossa tietueet on järjestetty jonkin ominaisuutensa, esimerkiksi tekijän sukunimen tai nimekkeen, perusteella. Haluttu tietue löytyy avaimen perusteella helposti ja tehokkaasti. Ylläpito on työläämpää kuin lineaarisen jonon, kun uudet dokumentit pitää lisätä oikeaan paikkaan, yleensä tiedoston keskelle. Indeksoidusta tiedostosta dokumentin löytäminen on helppoa. Järjestetyn peräkkäistiedoston jonkin ominaisuuden perusteella muodostetaan hakemisto eli indeksi. Esimerkiksi dokumenttiluettelossa kunkin kirjaimen kohdalle liitetään tieto siitä, minkä numeroinen dokumentti on ensimmäinen, jonka tekijän nimi alkaa tällä kirjaimella. Ylläpito vaatii työtä: aina kun kokoelmaan lisätään dokumentti, koko tiedosto ja indeksi täytyy päivittää. Käänteistiedosto on järjestetty avainten mukaisesti. Kuhunkin indeksiavaimeen on liitetty sitä käsittelevien dokumenttien numerot. Jo indeksin perusteella pystytään päättelemään, mitkä dokumentit täsmäävät kyselyyn. Useimmat kaupallisessa käytössä olevat hakujärjestelmät perustuvat käänteistiedostoihin. (Salton & McGill 1983, 12-18.)

2. Luonnollinen kieli ja tiedonhaku

Tekstitiedonhaun tulosjoukko syntyy kyselyn hakuavainten ja dokumenttien indeksointiavainten täsmäyttämisen perusteella. Jos dokumentit on kuvailtu formaalilla kielellä eli indeksointiavaimet on järjestetty kontrolloiduksi sanastoksi, hakuavaimet voi valita sanastosta ja ne täsmäytyvät hyvin käytettyihin indeksiavaimiin. Nykyisissä suurissa kokoteksti-indeksoiduissa tietokannoissa dokumentit on yleensä indeksoitu kaikilla niissä esiintyvillä sanoilla eli luonnollisella kielellä. Tämä lisää tiedonhaun ongelmia: Aihelueiden sanasto voi olla enemmän tai vähemmän vakiintunut. Sanat muodostavat hierarkioita. Useimmat kielet taipuvat, jotkut voimakkaastikin.

Formaalit kielet ovat artefakteja eli tarkoituksella luotuja. Niitä ovat esimerkiksi matematiikan, logiikan ja atk-ohjelmoinnin kielet sekä järjestetyt sanastot. Formaaleja kieliä on vaikea käyttää muihin tarkoituksiin kuin mihin ne on luotu. Niiden perussymbolien lukumäärä on pieni ja ilmausten merkitykset ovat täsmällisiä. (Karlsson 1998, 3.)

Luonnollisten kielten sanojen monimerkityksisyys ja merkitysten väliset rajat vaikeuttavat tiedonhakua. Kommunikaatiossa nämä ominaisuudet antavat kielille ilmaisuvoimaa ja joustavuutta: uusissakin tilanteissa pärjätään epätäsmällisellä luonnollisella kielellä, kun sanamerkitykset venyvät kattamaan uusiakin tapauksia. (Karlsson 1998, 3.) Luonnollisten kielten luokittelu jää tämän työn ulkopuolelle, koska luokittelu ei ole oleellinen tämän työn kannalta.

Luonnollista kieltä tutkii **kielitiede** eli **lingvistiikka**. Kieli ei ole yksi jakamaton järjestelmä vaan monikerroksinen, monimutkainen ja joustava järjestelmä. Kieli koostuu osajärjestelmistä, joiden välillä on monenlaisia suhteita. Kielen osajärjestelmä koostuu sille tyypillisistä perusyksiköistä ja niiden välisistä suhteista. Osajärjestelmiä on viisi: fonetiikka, fonologia, morfologia, syntaksi ja semantiikka. (Karlsson 1998, 15.)

2.1 Fonetiikka ja fonologia

Fonetiikan keskeisiä ongelmia ovat puheen tuottaminen ja tunnistaminen sekä puheen akustisen rakenteen analysoiminen. Fonetiikassa pyritään luomaan järjestelmä, jolla maailman kielten äänellisiä resursseja voidaan kuvata ja luokitella. (Karlsson 1998, 45).

Fonologia tutkii kielten äännejärjestelmiä. Tärkeä tutkimuskohde on foneettisten äänne-erojen tehtävien analyysi. Sillä pyritään määrittämään kielen rakenteen kannalta merkityksellisiä äänne-eroja. Ne muodostavat kielen äännejärjestelmän ytimen. Fonetiikka ja fonologia ovat hyvin lähellä toisiaan ja yhdessä ne muodostavat ehyen, konkreettisen tutkimusalueen. (Karlsson 1998, 63.) Fonetiikka ja fonologia ovat tärkeitä aloja puhetiedonhaussa, mutta koska tekstitiedonhaku ei hyödynnä fonetiikkaa ja fonologiaa lainkaan, näitä kielen osajärjestelmiä ei käsitellä tässä työssä enempää.

2.2 Morfologia

Morfologiassa eli muoto-opissa tutkitaan sanojen sisäistä rakennetta niiden morfologisten osasten eli morfeemien kannalta. Morfeemi on pienin merkityksenkantaja. Morfologian kytkee fonologiaan morfeemien fonologisen rakenteen ja sen vaihteluiden ehtojen selvittäminen. Morfeemien merkitysten ja tehtävien tutkiminen liittää morfologian läheisesti myös syntaksiin ja semantiikkaan. Morfologia koostuu taivutuksesta ja sananmuodostuksesta. (Karlsson 1998, 83.) Taivutusmorfologia tutkii esimerkiksi sijamuodoissa taipuneita sanoja (pilvissä). Sananmuodostusmorfologia tutkii kantasanoista syntyneitä uusia itsenäisiä sanoja, johdoksia (pilvetön) ja yhdyssanoja (pilvilinna). (Pirkola 2001.)

Taipuminen, johtuminen ja yhdyssanat vaikeuttavat tekstitiedonhakua. Vaikka sama indeksiavain esiintyy sekä dokumentin kuvauksessa että kyselyssä, dokumentti ei pääse tulosjoukkoon, elleivät indeksi- ja hakuavain ole samassa muodossa. Jos hakuavain on taivutetussa muodossa ja indeksi perusmuotoinen, muodot eivät kohtaa. Vähänkin taipuvissa kielissä indeksiavainten ja/tai hakuavainten morfologinen käsittely parantaa tiedonhaun tulosta. Käsittely voi olla sanan katkaisu, avainten stemmaaminen tai perusmuotoistaminen. (Pirkola 2001.)

Sana katkaistaan kohdasta, jossa taivutuksen vaikutus tai ehkä johdinkaan ei vielä näy. Jos sananvartalo katkaistaan liian aikaisin, hakutulokseen hyväksytään turhan erilaisia sanoja. Katkaisumerkki vaihtelee eri ohjelmissa, se voi olla esim. * tai $.

sana	**vartalo**	**sana**	**vartalo**
kaupungissa	kaupun*	epäkohtaa	epäkohta*,
kauppiaalle	kauppia*	epäkohtiin	epäkohti*
kauppaehdoin	kauppaeh*	epäkohtelias	epäkohteli*

Stemmaaminen on pääteaineksen karsimista avaimesta. Poistettavia elementtejä ovat liitteet, päätteet, tunnukset ja mahdollisesti johtimetkin. Stemmaamiseen on kehitetty myös verkossa toimivia ohjelmia. Yksi tunnetuimpia stemmausalgoritmeja on Porter Stemmer. Porter Stemmerin kehittäjä Martin Porter on tehnyt myös Snowball-stemmerin. Snowball-algoritmista on myös suomenkielinen kuvaus. Seuraavat stemmausesimerkit ovat suomenkielisestä Snowballista:

sana	**vartalo**	**sana**	**vartalo**
edeltäjien	edeltäj	innostu	innostu
edentäjiensä	edeltäjie	innostuessaan	innostue
edeltäjiään	edeltäjiä	innostuimme	innostui

Stemmaaminen eli sanan yhden taivutusvartalon tuottaminen on yleensä riittävä toimenpide esimerkiksi englanninkielisessä tiedonhaussa, sillä englannin sanat eivät taivu paljon. Suomenkielessä taas sanat taipuvat runsaasti ja yhdellä sanalla on usein monia taivutusvartaloita. Suomenkielisessä tiedonhaussa kannattaa tuottaa sanan kaikki mahdolliset taivutusvartalot. Esimerkiksi Lingsoftin Finstems-ohjelma tuottaa ne. Finstemsillä sain esimerkit:

sana	**vartalot**	**sana**	**vartalot**
yö	yö, öi	omena	omena, omeni, omenoi, omenoj
työ	työ, töi	käsi	käsi, kätt, käde, käte
luoda	luo, loi	tulla	tule, tuli, tull, tult, tulk

Perusmuotoistaminen on pääteaineksen karsimista sanasta niin, että sana jää perusmuotoon. Lingsoftin Fintwol-ohjelma kertoo annetun sanan muodon sekä sen perusmuodon. Fintwol tuotti esimerkit:

"<öisin>"
"öinen" A SUP NOM SG
"öinen" A POS INS PL
"öisin" ADV

"<tullessaan>"
"tulla" V INF2 ACT INE 3

```
"<käsittä>"
  "käsi"  N ABE PL
```

Jotkut hakujärjestelmät stemmaavat tai perusmuotoistavat avaimet automaattisesti hakijan puolesta. Etenkin voimakkaasti taipuvissa kielissä, kuten suomi, avainten morfologinen käsittely on erityisen hyödyllistä. (Pirkola 2001.) Alkula (2000, 151-152) toteaa väitöskirjassaan, että perusmuotoistaminen vähensi merkittävästi indeksiavainten määrää suomenkielisessä kokoelmassa. Tällä säästettiin selvästi tallennustilaa. Perusmuotohakemistosta tehdyn haun tarkkuus oli myös selvästi parempi kuin vastaavan taivutusmuotokyselyn tarkkuus. (Alkula 2000, 186.)

2.3 Syntaksi ja sanasto

Syntaksin eli lauseopin tutkimuskohteet ovat lauseiden rakenne sekä lauseen muodostavat pienemmät osat ja osien suhteet, tehtävät ja yhdistäminen. Lause koostuu sanoista. Sanat järjestyvät hierarkkisesti lausekkeiksi. Syntaksi näkyy muun muassa sanajärjestyksessä, taivutuksessa ja kongruenssissa. Lauseet ovat nimenomaan kirjoitetun kielen perusyksikkö, puhekielessä on enemmän epätäydellisiä lauseita. (Karlsson 1998, 120.) Syntaksia on hyödynnetty tekstitiedonhaussakin esimerkiksi kieltenvälisen tiedonhaun vastinkorpusten muodostamisessa (ks. esim. Nie, Simard, Isabelle & Durand 1999). Koska syntaksi on tämän työn kannalta varsin vähän merkityksellinen, en käsittele sitä tämän enempää.

Neljäs kielen osajärjestelmä on **sanasto**. Leksikaaliset sanat eli lekseemit ovat kielen sanaston ydin. Niiden tärkein ominaisuus on sanaluokka. Kielen ydinsanasto koostuu keskeisimmistä sanoista. Lisänä on tarpeellinen määrä erikoissanoja. Uusia sanoja muodostuu esimerkiksi uudismuodosteina, johtamalla, yhdistämällä, lainaamalla, lyhentämällä ym. Useimmat lekseemit ovat monimerkityksisiä. (Karlsson 1998, 186.)

2.4 Semantiikka

Lingvistinen **semantiikka** eli merkitysoppi tutkii luonnollisten kielten merkitysilmiöitä. Sen keskeinen ongelma on kielten sanojen ja kieliopillisten kategorioiden merkitykset kielijärjestelmän osina. Sananmerkitysten tutkimus on leksikaalista semantiikkaa. (Karlsson 1998, 200.) Leksikaalinen semantiikka on tämän työn kannalta tärkein kielitieteen osa-alue.

Luonnollisessa kielessä samalla tarkoitteella voi olla useampia ilmauksia ja yksi sanamuoto voi saada monia eri merkityksiä. **Polysemiasta** on kyse, kun samalla sanalla voidaan tarkoittaa useampaa asiaa. Esimerkiksi kieli on polyseeminen sana. Sillä on neljä merkitystä: suussa oleva elin; kieleke, läppä; soittimen osa; puheen järjestelmä. Polysemian takana on aina yhteinen juurisana, joka on saanut vakiintuneita erillisiä merkityksiä.

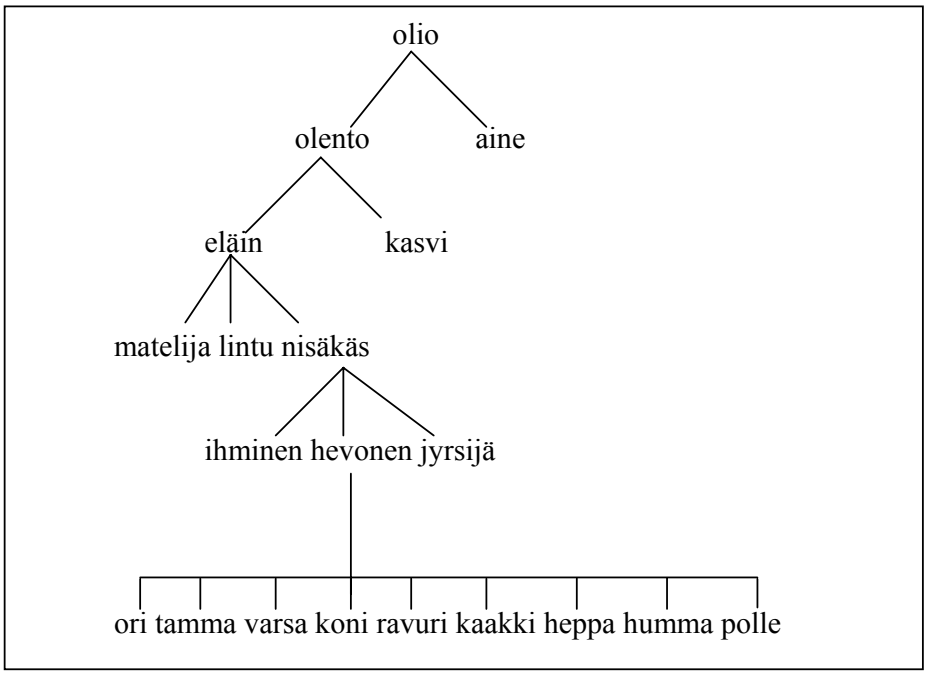
Polysemian kaltainen ilmiö on **homonymia**: äänneasultaan samalla sanalla on täysin eri merkityksiä. Homonyymit taipuvat eri tavalla ja niiden syntaktinen rooli voi olla erilainen. Esimerkiksi kuusi merkityksessä kuusipuu taipuu kuusi – kuusen – kuusta, kun taas kuusi merkityksessä 6 taipuu kuusi – kuuden – kuutta. Homonyymien taustalla ei ole yhteistä juurisanaa. (Karlsson 1998, 213.)

Synonymia on toiselta nimeltään samanmerkityksisyys. Synonymiassa kahdella tai useammalla muodoltaan eri sanalla on sama merkitys. Täydellisesti toisiaan vastaavia synonyymeja on harvassa, esimerkiksi kakara ~ penska, koskaan ~ milloinkaan. Usein merkitykset eroavat joko käyttötavaltaan, tyyliarvoltaan, konnotaatioltaan tai affektiiviselta merkitykseltään. Lainautuminen, sanaston aktiivinen kehittäminen tai terminologian vakiinnuttaminen voi synnyttää lähisynonyymejä, joilla on tyylillisiä tai tekstilajikohtaisia eroja: suola ~ natriumkloridi, Yleisradio ~ YLE, tervehtiä ~ moikata. A. Leinon ja P. Leinon synonyymisanastossa (1992, tässä Karlsson 1998, 220) annetaan juosta-verbille seuraavat lähisynonyymit, joista osa on deskriptiivisiä tarkoitetta jollakin tavalla kuvaavia:

> harppoa, hipsutella, hissutella, hölkyttää, hölköttää, jolkottaa, kiitää, kipaista, kipittää, kirmaista, köpittää, laukata, livistää, lönkytellä, lönkyttää, löntystää, nelistää, piipertää, pinkaista, pinkoa, porhaltaa, pyyhältää, ravata, sipsuttaa, vemputella, vouhottaa.

Tämä lista osoittaa, että sanojen väliset rajat ovat sumeat. Suomessa deskriptiivisluontoisten verbimuunnosten tuottaminen on runsasta. (Karlsson 1998, 219-220.)

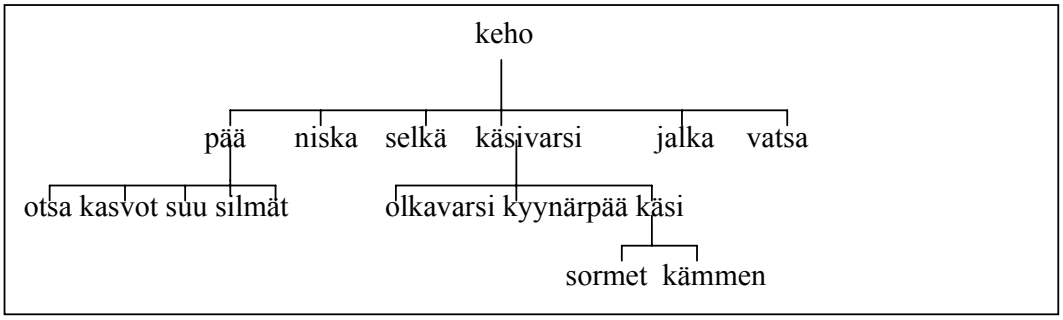
Hyponymialla kuvataan merkitysten hierarkkisia suhteita. Alisteinen sana on hyponyymi. Yläkategorian sana on hyponyymiin nähden hyperonyymi. Arkikielen sanat muodostavat usein merkityspiirteisiin perustuvia hierarkioita. Kuviossa 1 nisäkäs on sanan hevonen hyperonyymi. Hevonen on nisäkkään hyponyymi ja humman hyperonyymi. (Karlsson 1998, 221.)



Kuvio 1: Esimerkki arkikielen perussanojen hierarkiasta (Karlsson 1998, 221)

Meronymia (joskus myös partonymia) tarkoittaa osa—kokonaisuus -suhdetta. Arkielämässä moni asia koostuu luontaisista osista. Kuviossa 2 on kuvattu karkeasti kehon meronymiasuhteita. Meronymiassa osa, kuten pää, on meronyymi ja kokonaisuus, kuten keho, on holonyymi. (Karlsson 1998, 222.)

Synonymian vastakohta on **vastakohtaisuus**. Vastakohtia on useita lajeja. Lajivastakohtaa nimitetään komplementaariseksi vastakohdaksi. Komplementaarinen vastakohta on joko—tai -tilanne; muita vaihtoehtoja tai välimuotoja ei ole, esimerkiksi mies — nainen, naimaton —



Kuvio 2: Ylimalkainen kehon meronymiakaavio (Karlsson 1998, 222.)

naimisissa, elävä — kuollut. Lajivastakohtaisten adjektiivien erityispiirre on vertailuasteiden esiintymisen outous, esimerkiksi naimattomampi tai kuollein.

Toinen keskeinen vastakohtatyyppi on antonymia eli astevastakohtaisuus, esimerkiksi kuuma — kylmä, hyvä — paha, helppo — vaikea. Antonyymit ovat yleensä adjektiiveja, joilla kuvataan liukuvaa ominaisuutta. Niistä voi helposti muodostaa vertailuasteita mutta niiden tarkoitteista on vaikea muodostaa kahta täysin eroavaa ryhmää. (Karlsson 1998, 223.)

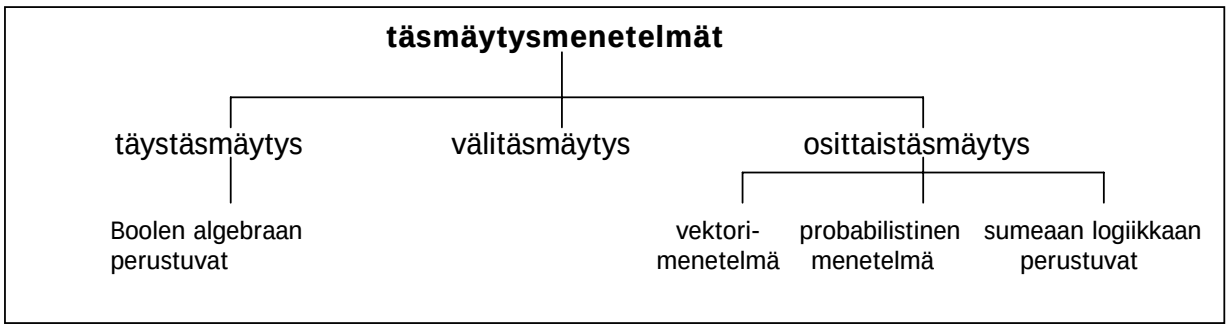
3. Hakujärjestelmät

3.1 Yleistä hakujärjestelmistä

Hakujärjestelmien perustehtävä on kyselyn hakuavainten ja dokumentin indeksointiavainten vertaaminen toisiinsa eli täsmäyttäminen. Täsmäytysmenetelmä selvittää, ovatko hakuavaimet ja indeksointiavaimet niin samanlaisia, että dokumentti kannattaa ottaa mukaan tulosjoukkoon. (Korfhage 1997, 79.)

Hakutilanteessa dokumentti ja kysely käyvät läpi samanlaisen tapahtumaketjun. Hakujärjestelmän ulkopuolella joku luo tai kerää dataa ja muodostaa siitä valmiin dokumentin. Jotta dokumenttia voidaan käsitellä täsmäytysprosessissa, siitä luodaan järjestelmään sisäinen esitys. Kyselykin saa alkunsa järjestelmän ulkopuolella, kun hakija tekee tiedon tarpeestaan kyselyn. Myös kyselystä tehdään järjestelmässä sisäinen esitysmuoto. Sellaisena sitä voidaan verrata dokumenttiin. Hyvä järjestelmä tekee muokkaukset yksinkertaisesti ja automaattisesti, käyttäjän huomaamatta. (Korfhage 1997, 51-52).

Täsmäytysmenetelmien pääjako on esitetty kuvassa 3. Täsmäytysmenetelmät arvioivat avainten samanlaisuutta eri perustein. **Täystäsmäyttävässä** (exact match) menetelmässä kaikkien hakuavainten tulee esiintyä kyselyssä esitetyssä muodossa myös dokumentin esityksessä, ja dokumentin esityksen on noudatettava hakulausekkeen avainten ja operaattoreiden logiikkaa. Täystäsmäytysmenetelmistä tunnetuin on Boolen algebraan perustuva menetelmä (Belkin & Croft 1987, 113). Täystäsmäytyksen laajennus on **välitäsmäytys** (range match). Haluttu hakuavain voidaan määrittää tietylle, esimerkiksi numeeriselle tai aakkoselliselle välille. (Korfhage 1997, 52). **Osittaistäsmäyttävissä** (partial match) menetelmissä hakutulokseen pääsee myös dokumentteja,



Kuvio 3: Täsmäytysmenetelmät (Korfhage 1997:n pohjalta)

jotka vastaavat kyselyä vain osittain. Avainten täytyy esiintyä samassa asussa sekä kyselyssä että dokumentin kuvailussa, mutta kaikkien hakuavainten ei välttämättä tarvitse esiintyä dokumentin kuvauksessa. Hakutulos voidaan järjestää relevanssijärjestykseen sen mukaan, mitkä dokumentit vastaavat parhaiten kyselyä. Yleisimpiä osittaistäsmäyttäviä menetelmiä ovat vektorimenetelmä, todennäköisyyslaskentaan perustuva eli probabilistinen menetelmä ja sumeaan logiikkaan perustuva menetelmä. (Korfhage 1997, 82-93.)

Tämän työn empiirisessä tutkimuksessa käytetään todennäköisyyslaskentaan perustuvaa eli probabilistista järjestelmää. Siksi sitä esitellään seuraavaksi tarkemmin ja muita täsmäytysmenetelmiä lyhyemmin. Sumean täsmäytyksen kuvaus jätetään kokonaan pois ja sillä tehtyjä tutkimuksia ei esitellä lainkaan.

3.2 Boolen menetelmä

Boolen hakujärjestelmässä kysely on annettujen sanojen looginen funktio (Korfhage 1997, 81). Kysely annetaan Boolen operaattoreilla muotoiltuna. Perinteiset Boolen operaattorit ovat AND, OR ja NOT. AND edellyttää, että kaikki sillä yhdistetyt hakuavaimet esiintyvät dokumentissa, OR taas, että vähintään yksi sillä yhdistetyistä avaimista esiintyy ja NOT, että yksikään sillä merkityn hakuavaimen sisältämä dokumentti ei saa tulla mukaan hakutulokseen. (Korfhage 1997, 53-54.) Operaattorin vaikutusalue kyselyssä osoitetaan sulkuja käyttämällä, esimerkiksi

> (shetlannin AND lammaskoira) OR (shetland AND sheepdog) OR shetlanninlammaskoira

Tutkimus on paljastanut täystäsmäyttävissä menetelmissä joitakin ongelmia. Puhdas Boolen menetelmä jättää löytämättä monia relevantteja dokumentteja, jotka täsmäävät vain osittain kyselyyn. Se ei myöskään järjestä dokumentteja relevanssin mukaiseen järjestykseen eikä ota huomioon kyselyn eikä dokumentin avainten tärkeyseroja. Boolen menetelmä on riippuvainen kahdesta esityksestä, jotka on muodostettu samasta sanastosta. (Belkin & Croft 1987, 113.)

Täystäsmäyttäviä menetelmiä on kehitetty täsmäyttämään vähemmän täydellisesti, ottamaan huomioon tärkeyseroja ja suorittamaan relevanssijärjestys. Hakusanan katkaisu tai korvausmerkin käyttäminen tuo joustoa täsmäytykseen. Tärkeyseroja ja relevanssijärjestys saadaan aikaan yhdistämällä osittais- ja täystäsmäyttäviä menetelmiä. (Belkin & Croft 1987, 114.)

3.3 Vektorimalli

Dokumentin sisältämiä sanoja voidaan pitää niiden ominaisuuksina tai piirteinä. Kullekin piirteelle eli indeksiavaimelle voidaan antaa paino sillä perusteella, kuinka keskeinen se on dokumentin sisällön kannalta ja kuinka hyvä erottelija se on kokoelmassa. Paino voi olla joko binäärinen (1, jos avain esiintyy dokumentissa ja 0 jos avain ei esiinny dokumentissa), tai ei-binäärinen, kun kunkin avaimen paino lasketaan esimerkiksi sen dokumentti- ja/tai kokoelmafrekvenssin perusteella. (Salton & McGill 1983, 201.)

Kahden dokumentin samankaltaisuus lasketaan yleensä dokumenttien yhteisten piirteiden funktiona. Jos avaimet on painotettu, laskenta voi perustua yhteisten avainten painoihin. Kun on kaksi dokumenttia DOC_1 ja DOC_2, dokumentin i indeksiavaimen k painoa kuvaa $TERM_{ik}$. Jos dokumentteja kuvataan niiden sisältämillä sanoilla eli piirteillä, voidaan esittää seuraavat piirrevektorit:

$DOC_1 = (TERM_{11}, TERM_{12}, \ldots, TERM_{1t})$

$DOC_2 = (TERM_{21}, TERM_{22}, \ldots, TERM_{2t})$

Vektorien samanlaisuuden laskentamenetelmiä on monia, esimerkiksi Dicen tai Jaccardin menetelmä, sekä tilastollisia menetelmiä (ks. esim. Salton & McGill 1983, 203). Yksi tunnetuimpia menetelmiä on kosinimenetelmä:

$$SIM\left(DOC_i, DOC_j = \frac{\sum_{k=1}^{t} \left(TERM_{ik},\ TERM_{jk} \right)}{\sqrt{\left(TERM_{ik} \right)^2 \left(TERM_{jk} \right)^2}} \right)$$

(Salton & McGill 1983, 201-204.)

3.5 Probabilistinen menetelmä

Probabilistinen tiedonhakujärjestelmä laskee todennäköisyyden, että dokumentti d on relevantti tietyn kyselyn q suhteen. Oletetaan, että kaikki kyselylle q relevantit dokumentit tunnetaan ja

niiden lukumäärä on n. Koko tietokannan kaikkien dokumenttien lukumäärä on N. Kun dokumentti d valitaan satunnaisesti tietokannasta, voidaan todennäköisyys, että d on relevantti kyselyn suhteen, laskea seuraavasti:

$$P(relevantti) = \frac{n}{N}.$$

Todennäköisyysteorian mukaan todennäköisyys, että dokumentti ei ole relevantti, voidaan laskea seuraavan kaavan mukaan:

$$P(\neg relevantti) = 1 - P(relevantti) = \frac{N-n}{N}.$$

Tulosjoukkoon valitaan ne dokumentit, jotka täsmäävät kyselyyn parhaiten. Tämä selvitetään vertaamalla kyselyn haku- ja dokumentin indeksointiavaimia. Kehittyneissä järjestelmissä arviointi voi perustua myös syntaksiin, semantiikkaan ja/tai pragmatiikkaan. (Korfhage 1997, 88-89.)

Haku jakaa dokumenttijoukon kahteen joukkoon, valittuihin ja ei-valittuihin. Kaikki valitut dokumentit eivät välttämättä ole relevantteja. Ehdollinen todennäköisyys, että valittu dokumentti on relevantti, merkitään:

$$P(relevantti|valittu).$$

Valitun dokumentin ei-relevanssin todennäköisyys merkitään:

$$P(\neg relevantti|valittu).$$

Jos dokumenttijoukko S on valittu kyselyn perusteella, S:n jokaiselle dokumentille pätee

$$P(relevantti|valittu) > P(\neg relevantti|valittu).$$

Koska näiden kahden todennäköisyyden summa on 1, saadaan

$$P(relevantti|valittu) > 0{,}5.$$

Tästä saadaan joukon lajittelufunktio

$$dis(valittu) = \frac{P(relevantti|valittu)}{P(\neg relevantti|valittu)}$$

eli hae alkio jos ja vain jos $dis(valittu) > 1$. (Korfhage 1997, 89-90.)

Koska dokumentin relevanssi arvioidaan sen sisältämien sanojen eli indeksiavainten perusteella, edelliset todennäköisyydet suhteutetaan dokumentin avainten esiintymismääriin. Todennäköisyysteoriaan pohjautuvan Bayesin teoreeman perusteella ehdollista todennäköisyyttä voidaan muokata seuraavalla tavalla:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Kun sijoitetaan tämä kaava aikaisempaan lajittelufunktioon, saadaan:

$$dis(valittu) = \frac{P(relevantti|valittu)}{P(\neg relevantti|valittu)} = \frac{P(valittu|relevantti)P(relevantti)}{P(valittu|\neg relevantti)P(\neg relevantti)}.$$

Osoittaja on sen tapahtuman todennäköisyys, että dokumentti valitaan, jos se on relevantti, kerrottuna todennäköisyydellä, että kokoelman satunnainen, nyt tarkasteltava, dokumentti on relevantti. Nimittäjä on ei-relevantin dokumentin vastaava tulo. (Korfhage 1997, 90-91)

Oletetaan, että dokumenttia kuvaa avainjoukko $t_1, \ldots, t_n$ ja että nämä avaimet ovat tilastollisesti toisistaan riippumattomia. Dokumentin todennäköisyys voidaan tämän perusteella laskea avainten todennäköisyyksien tulona:

$$P(valittu|relevantti) = P(t_1|relevantti)P(t_2|relevantti)...P(t_n|relevantti)$$

ja vastaavasti tapaukselle $P(valittu|\neg relevantti)$. (Korfhage 1997, 91.)

Kun tunnetaan eri avainten esiintymisen todennäköisyys relevanteissa ja ei-relevanteissa dokumenteissa, voidaan arvioida todennäköisyys, että dokumentti saadaan hakutulokseen ehdolla että se on relevantti tai ei-relevantti. Jos tiedetään satunnaisesti valitun dokumentin relevanssin todennäköisyys, voidaan laskea annetun dokumenttijoukon lajittelufunktio ja päättää, hyväksyäkö joukko kyselyn hakutulokseksi. (Korfhage 1997,91.)

3.6 Menetelmien vertailu

Belkin & Croftin (1987, 124) mukaan menetelmiä verrattaessa on saatu joitakin yhdenmukaisia tuloksia. He toteavat, että osittaistäsmäyttävät menetelmät toimivat paremmin kuin täystäsmäyttävät. Vaikka hakujen tulosjoukkojen vertailu on ongelmallista, heidän mukaansa tuloksellisuuden ero on selvä.

Vuonna 1987 tekemässään vertailussa Belkin & Croft totesivat, että probabilistinen menetelmä on kaikkein tuloksellisin, kun siihen sisältyy avainten painotus ja tf.idf—kosinikorrelaatio -yhdistelmä. Menetelmä käyttää yksinkertaista samanlaisuusfunktiota, sisäistä tuloa. Tehokkuus perustuu indeksiavainten painotukseen. (Belkin & Croft 1987, 124.) 15 vuoden aikana tiedonhakua on tutkittu paljon lisää. Korfhage (1997, 92) toteaa, että probabilistisella hakumenetelmällä on saatu hyviä tuloksia. Ne eivät kuitenkaan ole olleet niin paljon parempia kuin Boolen logiikkaan perustuvien menetelmien tulokset, että järjestelmien kehittäjät olisivat joukolla siirtyneet Boolesta probabilistiseen järjestelmään. Tulevaisuuden kehityksestä Korfhage arvelee, että tiedonhaku on muuttumassa internetitse heterogeenisista tietokannoista tehtävään kokotekstihakuun. Probabilistiset menetelmät yleistyvät, kun tiedonhaku muuttuu entistä monimutkaisemmaksi. (Korfhage 1997, 92.)

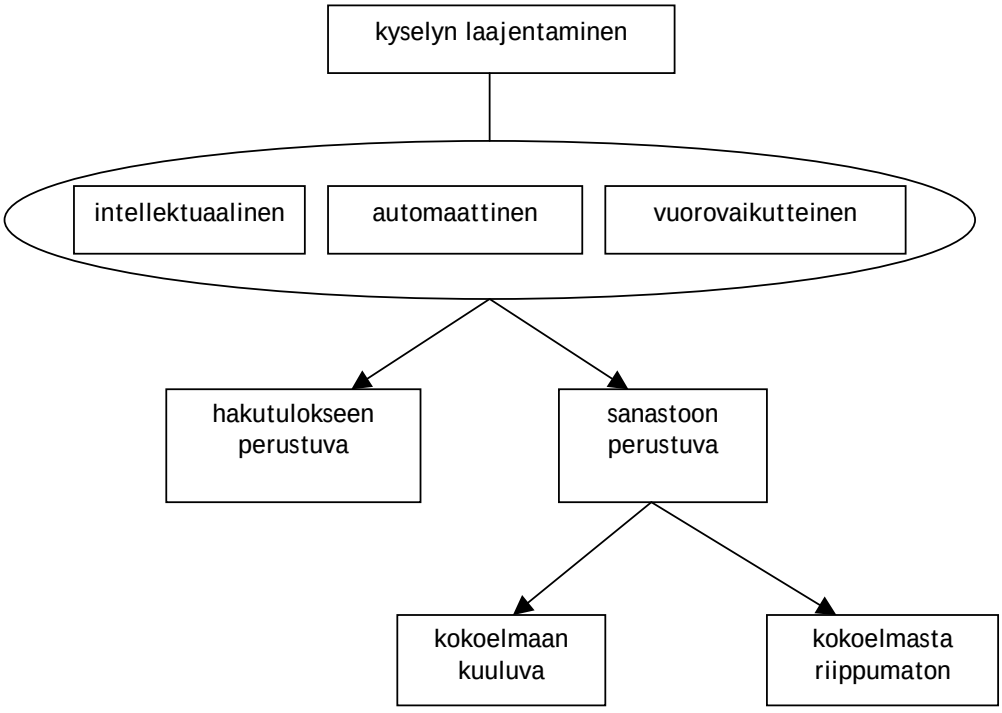
Vaikka eri hakumenetelmillä saadaan samantasoisia tuloksia, menetelmät löytävät usein samaan kyselyyn eri relevantit dokumentit. Poikkeavien sisällönkuvailumenetelmien, kuten viittaustietojen, käyttö voi myös johtaa eri dokumenttien löytymiseen. Menetelmien tuloksellisuus vaihtelee myös eri kyselyissä. Jos pystyttäisiin valitsemaan kullekin kyselylle juuri sen ominaisuuksilla parhaiten toimiva menetelmä, saataisiin aikaan upeasti toimiva järjestelmä. Ongelma on tietysti tunnistaa, mikä tekniikka on ihanteellinen millekin kyselylle. (Belkin & Croft 1987, 125-126.)

4. Kyselyn laajentaminen

4.1 Kyselyn laajentamisen vaihtoehdot

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 relevanssipalautteeseen, tai hakuprosessin ulkopuolisesta sanastosta, joka voi olla joko dokumenttikokoelmaan kuuluva tai siitä täysin riippumaton. Nämä vaihtoehdot on kuvattu kuviossa 4. (Efthimiadis 1996, 122-124.)

Intellektuaalisesti kyselyä laajentaa hakija itse valitsemansa hakustrategian ja –taktiikan perusteella. Hakija muotoilee itse alkuperäisen kyselyn ja lisää siihen itse valitsemansa laajennusavaimet. Intellektuaalinen laajentaminen on yleisintä käytettäessä Boolen menetelmää online-haussa ja CD-ROM-haussa. (Efthimiadis 1996, 126.)



Kuvio 4: Kyselyn laajentaminen: menetelmät ja termien lähteet (Efthimiadis 1996, 124).

Hyvä hakutulos edellyttää oikeiden hakutaktiikoiden valintaa. Osa taktiikoista voidaan valita ennalta hakustrategian perusteella. Muut taktiikat hakija valitsee ja toteuttaa oman harkintansa mukaan haun aikana. Haun edetessä syntyvän palautteen perusteella joitakin taktiikkoja muokataan tai hylätään. Alustavien tulosten perusteella hakijan tietämys muuttuu. Hän voi saada uuden näkökulman ja uutta ymmärrystä aiheeseen, mikä taas voi muokata hakuprosessia. Tuloksellisen hakutaktiikan valinta riippuu paljolti hakijan kokemuksesta ja arviointikyvystä. (Efthimiadis 1996, 131-132.)

Kun kyselyä laajennetaan **automaattisesti**, järjestelmä laajentaa alkuperäistä kyselyä. Efthimiadis (1996, 143-144) kertoo esimerkkinä automaattisesta kyselyn laajentamisesta mm. SMART-järjestelmästä, jossa kyselyn laajentaminen perustuu normalisoituihin vektoreihin. SMARTin laajennusavainten lähde on hakutuloksen dokumenttijoukko. Alkuperäiseen kyselyvektoriin lisätään joka muokkauskierroksella hakutuloksen relevanttien dokumenttien avainvektorit ja ei-relevanttien dokumenttien avainvektorit vähennetään. Kyselyyn saadaan lisää avaimia, joista osalla on negatiivinen paino. Myöhemmin menetelmää on kehitetty siten, että avain lisätään kyselyvektoriin, jos se esiintyy vähintään puolessa relevanteista dokumenteista ja lisäksi enemmän relevanteissa kuin ei-relevanteissa dokumenteissa.

Dokumenttikokoelman eli korpuksen sanoista voidaan myös rakentaa sanasto automaattisesti niiden yhteisesiintymisten perusteella. Korpuksessa useimmin yhdessä esiintyvistä sanoista muodostetaan ryhmiä. Kyselyavaimet laajennetaan ryhmänsä muilla sanoilla. Efthimiadis toteaa, että tutkimus ei ole osoittanut, että avainten yhteisesiintymiseen perustuva automaattinen kyselyn laajentaminen parantaisi merkittävästi hakutulosta. Kyselyavainten kokoelmafrekvenssi on yleensä korkea, joten myös niiden ryhmätovereiden kokoelmafrekvenssi on korkea. Kyselyyn tulee siis lisää yleisiä sanoja. Yleiset sanat ovat huonoja relevanttien ja ei-relevanttien dokumenttien erottelijoita, joten kyselyyn lisätyt avaimet tuskin parantavat hakutulosta. (Efthimiadis 1996, 147-151.)

Laajennusavaimet voidaan ottaa myös intellektuaalisesti rakennetusta, tietokannan ulkopuolisesta sanastosta. Ne ovat usein tesauruksia, joissa sanat on järjestetty hierarkkisesti. Kyselyä voidaan laajentaa hierarkiassa ylä-, alapuolisilla tai rinnakkaisilla termeillä, esimerkiksi synonyymeilla. Aihetta on tutkittu paljon, ja tulokset ovat olleet varsin ristiriitaisia. (Efthimiadis 1996, 154-156.) Oma empiirinen tutkimukseni on nimenomaan automaattisen laajentamisen tutkimista, joten perehdyn seuraavissa luvuissa joihinkin siitä aiemmin tehtyihin tutkimuksiin. Luvussa 4.3 esittelen tutkimuksen, jossa laajennusavainten lähde on tietokannasta riippumaton, intellektuaalisesti

rakennettu yleissanasto. Luvuissa 4.2 ja 4.4 laajennusavaimet otetaan kokoelmaa varten intellektuaalisesti räätälöidystä sanastosta. Luvussa 4.5 verrataan automaattista kyselyn laajentamista kolmella erilaisella sanastolla. Omassa kokeessani vertaan kahta laajennusavainten lähdetyyppiä: kaupallista, kokoelmasta riippumatonta yleissanastoa ja kokoelmaa varten räätälöityä sanastoa.

Kun kyselyä laajennetaan **vuorovaikutteisesti**, järjestelmä antaa käyttäjän valittavaksi hakutermejä uudelleenmuotoiluvaiheessa. Käyttäjä siis määrittää termien suhteellisen tärkeyden ja käytettävyyden. Haun onnistumisen tai epäonnistumisen syitä on siis vaikeampi päätellä, kun tulokseen vaikuttavia muuttujia on entistä enemmän ja niiden vaikutusta tulokseen ja toisiinsa on vaikea määrittää. Termien lähde voi olla, samoin kuin automaattisessa laajentamisessa, joko hakutuloksiin perustuva tai sanasto, joka voi olla joko itsenäinen tai kokoelmaan liittyvä. (Efthimiadis 1996, 156).

Magennis & Rijsbergen tutkivat kokemattomien tiedonhakijoiden vuorovaikutteista kyselyn laajentamista. Yleisesti ottaen hakijoiden valitsemat laajennustermit eivät parantaneet haun tulosta. Tutkijat toteavat, että kokemattomien hakijoiden heikko menestys kertoo, että kyselyn vuorovaikutteinen laajentaminen on vaikea tehtävä. Perinteinen relevanssiarvioon perustuva kyselyn laajentaminen on käyttäjälle helpompaa ja yksinkertaisempaa. Toisaalta kokeneelle käyttäjälle vuorovaikutteisuus suo enemmän vaikutusmahdollisuuksia kyselyn etenemisessä (Magennis & Rijsbergen, 1997).

Kyselyn laajentamista on tähän asti tutkittu enimmäkseen rakenteettomilla eli litteillä kyselyillä. Viime vuosina suoritetuissa tutkimuksissa on kuitenkin osoitettu, että laajentaminen hyödyttää rakenteisia kyselyjä enemmän kuin litteitä (Kekäläinen 1999, 87; Kekäläinen & Järvelin, 2000)

4.2 Kyselyn laajentaminen hakutesauruksen avulla Boolen logiikkaan perustuvassa järjestelmässä

Kristensen (1992) tutki hakutesauruksen käyttöä kyselyn laajentamisessa. Tutkimus tehtiin suomenkielisellä Aamulehden noin 225 000 artikkelia sisältävällä tietokannalla. Hakujärjestelmä oli BASIS, joka perustuu Boolen logiikkaan ja käänteistiedostorakenteeseen. Boolen operaattoreiden lisäksi järjestelmässä voi käyttää läheisyysoperaattoreita. Tietokannan käänteistiedoston avaimet perusmuotoistettiin MORFO-ohjelmalla. Kristensen rakensi hakutesauruksen itse, koska valmista,

sanomalehden aiheisiin ja kieleen sopivaa tesaurusta ei löytynyt. Tesauruksen aihealueeksi valittiin talous ja ympäristöasiat ja lähtökohdaksi sanomalehden näkökulma.

Hakuaiheet (yhteensä 30) saatiin Aamulehden toimittajilta ja tietopalveluyritykseltä. Kunkin kysymyksen esittänyt toimittaja arvioi laajimman haun antaman tulosjoukon jokaisen dokumentin relevanssin. Tietopalveluyritykseltä saatujen aiheiden relevanssin arvioi yrityksen työntekijä. Arviot olivat binäärisiä: relevantti tai ei-relevantti. Perushakuja laajennettiin 1) synonyymeillä, 2) suppeammilla termeillä, 3) rinnakkaistermeillä ja 4) kaikilla edellisillä ryhmillä (=laajin haku).

Näin suuresta tietokannasta absoluuttisen saannin laskeminen olisi ollut mahdotonta, joten saannin mittarina käytettiin suhteellista saantia. Laajimman haun saanti sai arvon 100 % ja muita tuloksia verrattiin siihen. Tarkkuus laskettiin relevanssiarvioiden perusteella. Tutkimuksen tarkoitus oli selvittää hakutesauruksen kokonaisvaikutus hakutulosten saantiin ja tarkkuuteen sekä eri avaintyyppien menestys laajennusavaimina. Myös muita hakutulokseen vaikuttavia tekijöitä tarkkailtiin: läheisyysoperaattorin käyttöä JA-operaattorin tilalla, yhdyssanojen jakamista osiinsa ja hakuavainten dokumenttifrekvenssiä tulosjoukossa. (Kristensen 1992, 20-27.) Tulosten tilastollista merkitsevyyttä mitattiin Wilcoxonin järjestyssummatestillä, kun verrattiin kahden menetelmän keskinäistä eroa, ja Friedmanin testillä, kun verrattiin viiden eri hakutyypin eroja. (Kristensen 1992, 32-33.) Koska vain kyselyn laajentaminen on olennaista oman työni kannalta, käsittelen Kristensenin tuloksia vain tältä osin.

Kaikilla ryhmillä laajentaminen kaksinkertaisti perushaun tuloksen suhteellisen saannin. Tarkkuus taas heikkeni noin kymmenen prosenttiyksikköä. Saanti parani tilastollisesti merkitsevästi ($p<0,0001$), mutta myös tarkkuuden huononeminen oli tilastollisesti merkitsevää ($p<0,01$). Synonyymeilla, suppeammilla termeillä ja rinnakkaistermeillä laajentamisen saannit ja tarkkuudet olivat varsin samanlaisia, mutta tulosjoukot sisälsivät verraten vähän samoja artikkeleita. Paras saanti saatiin kaikilla hakuavainvaihtoehdoilla laajentamalla.

Eri laajennusmenetelmistä synonyymi- ja rinnakkaistermilaajennosten tulosjoukoissa oli eniten yhteisiä artikkeleita. Suppeammilla termeillä laajentamalla saavutettiin paras tarkkuus, mutta suhteellinen saanti oli muita laajennusmenetelmiä heikompi. Rinnakkaistermihaun tarkkuus ei ollut merkitsevästi ($p<0,5$) parempi kuin laajimman haun tarkkuus. Suhteelliseen saantiin ei myöskään tullut merkitsevää eroa. Sekä synonyymi- että suppeampi termi -laajennuksen suhteellisen saannin ero laajimpaan hakuun on merkitsevä ($p<0,05$).

Kristensen toteaa, että vaikka rinnakkaistermihaku näyttäisi toimivan laajinta hakua paremmin, merkitsevyystaso oli pieni ja rinnakkaistermihaun ja synonyymi- ja suppeampi termi -haun välillä ei ollut eroja. Hänen mukaansa rinnakkaistermeillä laajentamista ei voi pitää ehdottomasti muita laajennuksia parempana. (Kristensen 1992, 46-48.) Kristensenin tutkimuksen mukaan siis synonyymilaajennus ei siis ole paras mahdollinen kyselyn laajentamistekniikka ainakaan Boolen logiikkaan perustuvassa järjestelmässä.

4.3 Kyselyn laajentaminen sanojen välisten semanttisten suhteiden perusteella vektorimalliin perustuvassa järjestelmässä

Voorhees (1994) tutki kyselyn laajentamista intellektuaalisesti leksikaalisten ja semanttisten suhteiden perusteella suuressa ja monialaisessa tekstikokoelmassa. Hän käytti laajennustermien lähteenä WordNet-sanastoa. WordNet on laaja yleistesaurus, jossa sanat on järjestetty taksonomioiksi. Yhtä käsitettä voi edustaa synonyymiryhmä. Tesauruksen neljä taksonomiaa perustuvat sanaluokkiin ja niille on määritetty useita suhteita, esimerkiksi suppeampi termi, laajempi termi tai rinnakkaistermi. (Mandala, Tokunaga, Tanaka, 2000.)

Testikokoelmana käytettiin TREC-kokoelmaa. Se on testikokoelma, joka syntyy tiedonhaun tutkimuksen TREC- ja Tipster-testausten tuloksena. Kokeessa käytettiin kokoelmasta osaa, joka sisälsi 742 000 englanninkielistä dokumenttia sanomalehdistä, teknisten kirjoitusten abstrakteista ja Federal Registeristä. Kyselyjä tehtiin 50 ja relevanssiarviot tuotettiin TREC-2 ja Tipster-3 -evaluoinneista. Laajennustermit valittiin intellektuaalisesti sopimattomien laajennustermien välttämiseksi. Testin hakujärjestelmä oli vektorimalliin perustuva SMART. (Voorhees, 1994.)

TREC-aiheen kuvaukseen lisättiin kenttä, joka sisälsi manuaalisesti WordNetistä valitut synonyymiryhmät. Kyselyn kentät indeksoitiin SMART-järjestelmän indeksointitoiminnolla. Kysely laajennettiin lisäämällä synonyymit kyselyvektoriin. Kyselyvektorit koostuivat alavektoreista, jotka olivat eri käsitetyyppejä (ctype). Kukin käsitetyyppi edusti erilaista leksikaalista suhdetta. Tyyppejä oli yksitoista: alkuperäiset kyselytermit, synonyymit ja kaikki WordNetin suhdeluokat. (Voorhees, 1994).

Dokumenttivektorin D ja kyselyvektorin Q samanlaisuus määritettiin D:n ja kunkin kyselyn alavektorin samanlaisuuden summana:

$$sim(D,Q) = \sum_{ctype\ i} \alpha_i \times D \bullet Q_i$$

missä '●' merkitsee kahden vektorin pistetuloa, Q_i on Q:n i:s alavektori ja α_i on reaaliluku, joka kuvaa käsitetyypin i tärkeyttä suhteessa muihin tyyppeihin. (Voorhees, 1994.)

Kokeessa laajennusavaimiksi valittiin kaikki avainryhmät, jotka liittyivät sanastossa suoraan kyselyavaimeen. Laajennusavainten tasot olivat synonyymit, alemmat ja ylemmät termit. Laajentamisen vaikutus hakutulokseen oli selvästi heikko. Kyselyjä oli kolmen pituisia. Pisimmän kyselyn avaimet otettiin koko aiheen kuvauksesta. Tämän kyselytyypin keskipituus oli 52,54 avainta. Toiseksi pisin kyselytyyppi sai avaimensa yhteenvetokentästä ja aiheen kuvauksessa annetuista avainkäsitteistä. Näiden kyselyjen keskipituus oli 29,22 sanaa. Lyhin kyselytyyppi muodostui pelkästä yhteenvetokentästä. Sen keskipituus oli 11,02 sanaa. Lyhin testattu kyselytyyppi oli ainoa, jossa laajentaminen paransi hakutulosta merkitsevästi: 35 % 11 pisteen tarkkuuksien keskiarvoa. Voorhees testasi myös laajennustermien automaattista valintaa lyhimmällä kyselytyypillä ja erilaisilla α:n arvoilla laskettuna. Mikään laajennetun kyselyn tulos ei ollut merkittävästi parempi kuin alkuperäisen kyselyn tulos. (Voorhees 1994.)

Voorheesin (1994) mukaan kyselyn laajentaminen on saantia parantava menetelmä, joten hänen mukaansa ei ole yllättävää että laajentaminen ei hyödytä pitkiä kyselyjä yhtä paljon kuin lyhyitä. Koska tiedonhakijat eivät yleensä laadi yksityiskohtaisia kyselyjä, sanojen välisiin leksikaalisiin ja semanttisiin suhteisiin perustuva kyselyn laajentaminen voi parantaa selvästi alkuperäistä kyselyä.

4.4 Kyselyn kompleksisuuden, laajentamisen ja rakenteen vaikutus probabilistisella järjestelmällä

Kekäläinen (1999) tutki todennäköisyyteen perustuvalla tiedonhakujärjestelmällä fasettipohjaisen kyselyn kompleksisuuden, laajentamisen ja rakenteen vaikutuksia hakutulokseen. Tutkimukset suoritettiin probabilistisella InQuery-järjestelmällä suomenkielisessä TUTK-tietokannassa. TUTK sisältää lähes 54 000 artikkelia sanomalehdistä ja kattaa talous-, ulkomaan ja kotimaan uutisia. Kokoelma sisältää 35 aihetta. (Kekäläinen 1999, 57-58.)

Kyselyt tehtiin intellektuaalisesti käsitefaseteista. Menetelmässä tunnistetaan ensin käsitetasolla hakupyynnöstä hakukäsitteet eli muodostetaan käsitefasetit. Seuraavaksi lingvistisellä tasolla käsitteet yhdistetään ilmaisuihin eli fasetin sisältämiin sanoihin ja sanayhdistelmiin.

Merkkijonotasolla ilmaisuista saadaan hakuavaimia, joita voidaan verrata tietokannan indeksiavaimiin, ja tehdystä kyselystä tulee hakujärjestelmälle oikealla tavalla muotoiltu kysely. Kyselyissä käsitefasetit jaetaan pääfaseteiksi ja apufaseteiksi sen mukaan, kuinka paljon käsitteen mukanaolo tai pois jättäminen vaikuttaa kyselyn tulokseen. (Kekäläinen 1999, 32-36.)

Tesauruksessa on käsitteitä, ilmauksia ja täsmäytysmalleja sekä näiden kohteiden välisiä suhteita. Käsitteitä on yhteensä 832. Käsitteellä voi olla useampia ilmauksia, joita tesauruksessa on yhteensä 1345. Täsmäytysmalli kuvaa, kuinka ilmaisu voidaan täsmäyttää dokumenttiavaimiin kyselykielestä riippumatta. Täsmäytysmalleja on yhteensä 1558. Käsitteiden väliset suhteet ovat joko rinnakkaistermi- tai hierarkkisia suhteita. Synonyymisuhteita ei esiinny käsitteiden, vaan käsitteen ilmausten, välillä. (Kekäläinen 1999, 59-67.)

Tesaurus toimii yhdessä ExpansionTool-ohjelmiston kanssa. ExpansionTool on kyselyn muodostamis- ja laajennusväline. Se on tarkoitettu käytettäväksi luonnollisen kielen tekstihaussa ennen haun suorittamista heterogeenisessä dokumenttikokoelmassa, jota ei ole indeksoitu intellektuaalisesti. Se tukee kyselyn automaattista muodostamista ja laajentamista. Se hallitsee erilaisia kyselyrakenteita ja vaihtelevia muita laajennusparametreja. Hakija voi laajentaa kyselyä, vaikka ei tunne kyselyjen rakenteita tai niiden vaikutusta laajentamiseen. Käsitteellisen laajentamisen tekee operaattori sanojen välisten suhteiden perusteella syklisessä käsiteverkostossa. (Järvelin, Kekäläinen, Niemi, 2001.)

Kekäläisen tutkimus oli laaja ja siinä tutkittiin mainittujen tekijöiden vaikutusta hakutulokseen hyvin monipuolisesti. Merkittävimpiin löytöihin kuuluu kyselyn rakenteen tärkeys eri laajuisissa kyselyissä. Vain lyhyissä kyselyissä rakenne ei vaikuta merkittävästi hakutulokseen. Toinen merkittävä tulos on käsitepohjaisen kyselyn tesaurustermeillä laajentamisen hyvä menestys. Parhaaseen tulokseen päästään laajentamalla sopivilla operaattoreilla muotoiltuja fasettimuotoisia kyselyjä. Rakenteettomissa kyselyissä laajentamisen vaikutus on jopa negatiivinen tai merkityksetön, kun taas rakenteisissa SYN-kyselyissä hakutulos paranee merkitsevästi. Paras tarkkuus saavutettiin suurimmalla laajennuksella eli lisäämällä synonyymit, suppeammat käsitteet ja rinnakkaiskäsitteet. (Kekäläinen 1999, 128.) Kekäläinen arvelee, että hyödyllinen kyselynlaajennusstrategia voisi olla pääkäsitteistä muodostettujen fasettien laajentaminen suppeammilla käsitteillä ja sivukäsitteistä muodostettujen fasettien laajentaminen rinnakkaiskäsitteillä (Kekäläinen 1999, 131).

Kekäläinen tutki myös painotusmenetelmien vaikutusta hakutulokseen. Sekä hakuavaimia että fasetteja painotettiin. Ensimmäisessä menetelmässä alkuperäisen kyselyn hakuavaimet saivat suuremman painon kuin laajennusavaimet. Toisessa menetelmässä pääfasetit saivat suuremman painon kuin sivufasetit. Alkuperäisten hakuavainten painottamisesta ei ollut suurempaa hyötyä. Pääfasettien painottaminen taas osoittautui hyödylliseksi strategiaksi. (Kekäläinen 1999, 133.)

Järvelinin ja Kekäläisen tutkimus (2000) vahvisti Kekäläisen tuloksen, että kyselyn laajentaminen hyödyttää enemmän vahvasti rakenteisia kyselyjä. Tutkimus suoritettiin samoin InQueryllä TUTK-tietokannassa laajentamalla käsiteperustaisia kyselyjä semanttisten suhteiden perusteella. Tutkimuksen tarkoitus oli selvittää relevanssin asteen merkitystä tiedonhakujärjestelmien evaluoinnissa. Dokumenteilla osoitettiin olevan erilaisia ominaisuuksia eri relevanssitasoilla. Dokumentin ominaisuuksien tilastollisten erojen osoitettiin voivan selittää hyvää tulosta käsitepohjaisessa kyselyn laajentamisessa. Eri relevanssitasoilla testattiin eri rakenteisia ja eri lailla painotettuja kyselyjä. Eri kyselytyyppien välillä havaittiin pienemmät suorituserot vain vähän relevanttien dokumenttien joukossa kuin relevanttien dokumenttien joukossa. Laajentamisen todettiin parantavan vahvasti rakenteisten kyselyjen tulosta enemmän kuin litteiden. (Järvelin & Kekäläinen, 2000.)

Relevanssia parantavia dokumentin piirteitä ovat esimerkiksi aiheen runsas käsittely, paljon kyselyn aiheeseen liittyviä sanoja ja aiheen käsittely monelta. Siis voimakkaasti rakenteiset laajennetut kyselyt, joissa esiintyy useita kyselyn eri aspekteihin viittaavia hakuavaimia, löytävät relevanteimmat dokumentit. (Järvelin & Kekäläinen, 2000.)

4.5 Vertailu: laajennustermit eri lähteistä

Mandala, Tokunaga ja Tanaka (2000) vertasivat erilaisten tesaurusten vaikutusta samojen kyselyjen laajentamiseen. Testiympäristönä heillä oli TREC-7 -kokoelma, joka sisältää 528 155 dokumenttia 50 aihepiiristä. Artikkelit ovat peräisin seuraavista lähteistä: Financial Times, Federal Register, Foreign Broadcast Information Service ja LA Times. Hakujärjestelmänä oli SMART, vektorimalliin perustuva hakujärjestelmä, jossa avainten painot lasketaan termifrekvenssin ja käänteisen dokumenttifrekvenssin perusteella ja tulos normalisoidaan dokumentin pituudella.

Vertailtavat tesaurukset olivat

- WordNet (intellektuaalisesti muodostettu tesaurus, kuvailtu luvussa 4.3),
- korpuspohjainen tesaurus, joka perustuu sanojen yhteisesiintymiseen ja niistä tilastollisesti määritettyihin samanlaisuuksiin ja näiden perusteella niputettuihin sanaryppäisiin.
- predikaatti—argumentti -pohjainen tesaurus, jossa suhteet muodostetaan lingvistisin perustein. Samanlaisessa kieliopillisessa kontekstissa esiintyvät sanat oletetaan samankaltaisiksi ja ne luokitetaan samaan luokkaan. Sanaluokkien muodostamiseksi määritettiin subjekti—verbi-, verbi—objekti- ja adjektiivi—nomini -suhteita.

Kokeissa kunkin kyselyn laajentamiseen käytettiin joko vain yhtä, kahta tai kaikkia kolmea tesaurusta. Yksinään käytettynä kaikki kolme tesaurusta tarjosivat hyvin samanlaista avainmäärää kyselyn laajentamiseen. Tesaurusten toiminnassa on selviä eroja: Korpuspohjainen tesaurus lisää uusien avainten lisäksi yhteisesiintymisen perusteella uusia avainsuhteita, joita WordNet ei löydä. Toisaalta automaattisesti muodostetun tesauruksen avainrakenteissa saattaa olla aukkoja.

Tutkimus osoitti, että tesaurusten erot täydentävät toisiaan ja kyselyn laajentaminen parantaa tulosta eniten, kun kaikkia kolmea käytetään yhdessä. Hakuavain, jota kaikki menetelmät tarjoavat, saa suurimman painon, kun taas vain yhden tesauruksen tarjoama avain saa pienemmän painon. Näin voidaan karsia huonolaatuisten hakuavainten vaikutusta tulokseen. (Mandala, Tokunaga, Tanaka, 2000.)

4.6 Johtopäätökset aiemmasta tutkimuksesta

Käsiteltyjen tutkimusten tulokset kyselyn laajentamisesta olivat toisaalta selvästi rohkaisevia, toisten kokeiden tulokset jäivät heikoiksi. Kristensenillä (1992) suurimman parannuksen toi mahdollisimman voimakas laajennus. Yhdellä assosiatiivisella suhteella laajentaessa rinnakkaistermilaajennus oli tehokkain. Voorheesin (1994) tutkimuksessa ainoa merkittävä parannus saatiin lyhyiden kyselyjen tarkkuuteen synonyymeilla laajentamalla. Voorheesin tutkimuksessa laajemmat laajennusmenetelmät kuin pelkät synonyymit eivät parantaneet tulosta merkittävästi. Kekäläisen (1999) paras tulos saavutettiin laajentamalla yhdessä synonyymeilla, suppeammilla, laajemmilla ja rinnakkaistermeillä. Kekäläinen arvelee, että paras tulos saavutettaisiin pääfasetteja suppeammilla ja sivufasetteja rinnakkaistermeillä laajentaen.

Kristensenin ja Voorheesin kyselyt olivat ilmeisesti rakenteettomia. Kekäläinen puolestaan saavutti käsitepohjaisessa kyselyn laajentamisessa parhaan tuloksen rakenteisilla kyselyillä ja laajentamalla mahdollisimman voimakkaasti. Suurin merkitys oli siis kyselyn rakenteella, ei lyhyydellä. Tosin Kekäläisen kyselyt olivat saman pituisia kuin Voorheesin lyhyet kyselyt.

Mandala, Tokunaga ja Tanaka (2000) vertasivat keskenään ja yhteiskäytössä eri tesaurustyyppejä: intellektuaalisesti muodostettua, korpuksesta tilastollisesti muodostettua ja korpuksesta lingvistisesti muodostettua. Yksistään käytettynä kaikki tesaurukset toimivat yllättävän samalla tavalla, mutta paras tulos saavutettiin käyttämällä kaikkia kolmea tesaurusta yhdessä, sillä ne tukivat toistensa toimintaa.

5. Kyselyn automaattinen laajentaminen synonyymeilla probabilistisessa hakujärjestelmässä

Tein kokeeni suomenkielisessä TUTK-tietokannassa (TUTKissa) probabilistisella InQuery-hakujärjestelmällä. Laadin viisi erilaista kyselyä kolmestakymmenestä TUTKin kaikkiaan 35 aiheesta. Laajennusavaimet otin joko Finthes-synonyymisanastosta tai TUTKia varten räätälöidystä tesauruksesta. Finthesin laajennusavaimet perusmuotoistin Fintwolilla. Tuloksia arvioin saanti—tarkkuus -taulukoiden ja -käyrien avulla, 11 pisteen tarkkuuksien keskiarvoilla sekä kumuloituneella hyödyllä ja alennetulla kumuloituneella hyödyllä. Tulosten tilastollista merkitsevyyttä arvioin Friedmanin testillä.

Internet ja kaupalliset tietokannat tuovat tiedonhakujärjestelmät lähemmäs naiivia tiedonhakijaa, jolla ei ole taitoa muodostaa rakenteisia kyselyjä eikä aikaa perehtyä tiedonhaun menetelmiin tai järjestelmän käyttöohjeeseen. Halusin tutkia, auttaisiko automaattinen synonyymilaajennus tavallista ihmistä löytämään relevanttia informaatiota. 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.

Nykyisissä suurissa tietokannoissa ja Internetissä tulosjoukko on usein niin suuri, ettei hakija jaksa selata kuin murto-osan alkupään dokumenteista. Sen vuoksi järjestelmiä on tärkeää kehittää löytämään relevanteimmat dokumentit ja asettamaan ne tulosjoukon alkupäähän. Tein kokeeni kolmella eri tasoisella relevanssikorpuksella: 1) kaikki relevantit dokumentit, 2) melko ja erittäin relevantit dokumentit ja 3) vain erittäin relevantit dokumentit. Halusin selvittää, onko menetelmien välillä eroja eri relevanssitasoilla.

5.1 TUTK-kokoelma

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 1994, 64.) . Tietokannan juttujen pituuden keskiarvo on 233 sanaa (Kekäläinen 1999, 57).

TUTK-kokoelmassa on 35 hakuaihetta. Kekäläinen (1999, 58-59) karsi tutkimuksessaan aiheet 30:een aiheen laajennettavuuden perusteella: pois jätettiin aiheet, joissa laajentaminen ei parantanut tulosta. Omassa työssäni käytän näitä 30 aihetta. Aiheet on lueteltu liitteessä 1.

Sormunen (1994) on luokitellut hakuaiheet niissä esiintyvien käsitteiden perusteella. Hän toteaa, että hakutulokseen vaikuttaa muiden seikkojen ohessa se, kuinka rajaavia käsitteitä hakukysymys sisältää. Esimerkiksi henkilön nimi, maantieteellinen alue ja organisaation nimi rajaavat varsin hyvin kyselyn tulosjoukkoa. Näistä aiheista kirjoitettaessa käsitteen nimi mainitaan aina. Kun yksilökäsitteen tarkoitteesta poistetaan yksilölliset piirteet, puhutaan yleistarkoitteesta. Sormunen jakaa hakuaiheet niissä esiintyvien pääkäsitteiden perusteella neljään luokkaan:

1. Yleiskäsitehaut
   a) aihehaut
   b) maantieteellisesti rajatut aihehaut

2. Yksilökäsitehaut
   c) organisaationimihaut
   d) henkilönimihaut

(Sormunen 1994, 38-39.)

TUTKin dokumenttien relevanssi on arvioitu neliportaisella asteikolla. Arvioijat olivat kaksi informaatikkoa ja kaksi toimittajaa. Alkuperäisestä 35 aiheesta 19 oli kahden arvioijan arvioimia. Rinnakkaisista arvioista 73 prosenttia oli samoja. (Sormunen 1994, 72.) Kristensenin (1996) tutkimusta varten arvioitiin uudestaan hakutuloksen ne dokumentit, jotka eivät kuuluneet saantikantaan. Arvioijina toimi kolme alkuperäisen relevanssiarvion suorittajaa. Kutakin hakuaihetta arvioi 1-3 henkilöä. Niissä tapauksissa, kun samaa aihetta arvioi kolme henkilöä, yksimielisyys oli keskimäärin 88,9 prosenttia. (Kristensen 1996, tässä Kekäläinen 1999.)

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

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

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.

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 kolmen tasoisilla saantikannoilla. Tasolla kaikki relevantit relevanssikorpuksessa ovat mukana kaikki vähänkin relevantit dokumentit, tasot 1-3. Tasolla relevantit relevanssikorpukseen kuuluvat sekä melko että erittäin relevantit dokumentit tasoilta 2 ja 3. Tasolla erittäin relevantit tulosjoukkoon hyväksyttiin tason nimen mukaan vain erittäin relevantit eli tason 3 dokumentit.

TUTKin saantikantojen koko eli kullekin aiheelle relevanttien dokumenttien lukumäärä vaihtelee aineistossa suuresti. Liitteestä 3 käy ilmi, että 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 ja aiheisiin 17 (matka), 29 (ydinv) ja 30 (vihr) kaksi kappaletta. Relevanttien dokumenttien vähyys vaikuttaa selvästi esimerkiksi keskiarvotarkkuuksiin. Kun relevantteja dokumentteja on vain muutama, yhden löytyminen tai löytymättä jääminen tekee kymmenien prosenttien eron tarkkuuteen.

5.2 InQuery-hakujärjestelmä

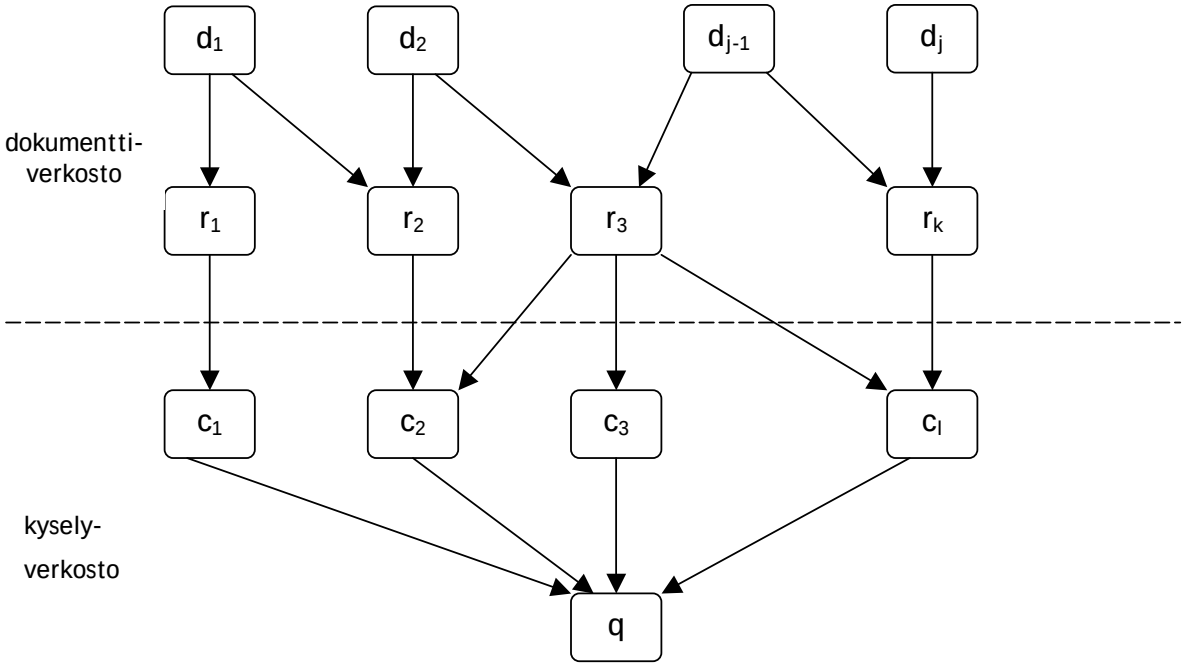
InQuery on todennäköisyyslaskentaan perustuva (probabilistinen), osittaistäsmäyttävä tiedonhakujärjestelmä. Se kehitettiin University of Massachusettsissa vastaamaan vaativiin tiedontarpeisiin suurissa tekstitietokannoissa. (Callan, Croft & Harding, 1992.)

InQuery perustuu probabilistiseen hakumalliin, päättelyverkkoon. InQueryn päättelyverkko on tyypiltään Bayesin verkko, jossa solmut esittävät propositiomuuttujia ja kaaret riippuvuuksia (ks. kuvio 5). Solmun arvo on joko tosi tai epätosi ja määräytyy sitä edeltävien solmujen arvojen perusteella. Kaarien arvot ovat välillä [0,1]. Järjestelmä koostuu kahdesta osaverkosta: dokumenttiverkko ja kyselyverkko. (Callan, Croft & Harding, 1992.)

Haku toimii siten, että dokumenttitekstin esitysmuotoa verrataan kyselyn esitysmuotoon niiden kielitieteellisten ja tilastollisten piirteiden perusteella. Tekstin esitysmuoto voi olla esimerkiksi sanoja, fraaseja, tekstikappaleita tai manuaalisesti annettuja avainsanoja. Kysely voi olla joko luonnollista kieltä tai operaattorein muodostettu hakulause. Käyttäjän antamaa kyselyä muotoillaan oppivilla tekniikoilla, kuten relevanssipalauteella ja reitittämällä. (Broglio, Callan & Croft, 1994).

Kuviossa 5 nähdään yksinkertainen dokumenttiverkosto. Verkostossa dokumentti kuvataan kahdella abstraktiotasolla: dokumentin tekstitasolla d ja sitä esittävien sanojen tasolla r. Tiedonhaussa dokumenttisolmu d_i kuvaa tilannetta, että dokumentti vastaa hakijan kyselyä. Dokumenttisolmu saa tällöin arvon tosi. Dokumenttisolmun d_i ja sisällön esityssolmun r_k välisen kaaren arvo on ehdollinen todennäköisyys $P(r_k|d_i)$. Ennakkotodennäköisyys $P(d_i)$ on $1/$(dokumenttien lukumäärä). (Callan, Croft & Harding, 1992.)

Kyselyverkosto kuvaa tiedontarvetta. Kuviossa 5 kyselyverkosto kuvataan kahdella abstraktiotasolla: kyselytaso q ja käsitetaso c. Kyselysolmu kuvaa tilannetta, että tiedontarve on tyydytetty. Kyselysolmun arvo on aina tosi. Käsitesolmu kuvaa tilannetta että käsite on havaittu dokumentissa. Käsitesolmun arvo voi olla joko tosi tai epätosi. Kyselyverkosto ja dokumenttiverkosto liittyvät toisiinsa käsitesolmujen ja sisällön esityssolmujen välityksellä. (Callan, Croft & Harding, 1992).



Kuvio 5: Yksinkertainen dokumenttihaun päättelyverkosto. (Callan, Croft, Harding, 1992)

Dokumenttiverkosto luodaan automaattisesti yhdistämällä dokumentit sisällön esityssolmuihin ja säilyttämällä solmutiedot käänteistiedostossa. Kyselyverkon tekee tiedonhakija. Dokumentin haku tapahtuu rekursiivisella päättelyllä laskemalla uskomusarvot päättelyverkostosta, ja sitten hakemalla parhaiksi arvioidut dokumentit. (Callan, Croft, Harding, 1992).

InQueryn päätoiminnot ovat dokumentin indeksointi, kyselyn käsittely, kyselyn evaluointi ja relevanssipalaute. **Dokumentin indeksoinnissa** dokumentit jäsennetään ja dokumenttien sisältöä kuvaavat indeksiavaimet tunnistetaan. Indeksoitavat osat tunnistetaan dokumentin rakenteen perusteella. Sanojen määrää vähennetään yleensä perusmuotoistamalla tai stemmaamalla. Perusmuodoista rakennetaan tiivistetyt käänteistiedostot. (Broglio, Callan, Croft, 1994.)

Kyselyn käsittelyprosessi tunnistaa keskeiset avaimet ja rakenteet, joita tiedonhakija on käyttänyt kuvaamaan tiedontarvettaan. Avainten suhteellinen tärkeys ja niiden väliset suhteet arvioidaan syntaksin perusteella. Avainten välisiä suhteita voi käyttää kyselyn laajentamiseen. Suhteet voivat perustua joko intellektuaalisesti laadittuun tesaurukseen tai korpusanalyysiin. InQueryn laajennusavainten löytämisväline on nimeltään WordFinder. Se etsii tekstistä nominiryppäitä ja paikantaa hakuavaimiin läheisesti liittyviä sanoja, ts. sanoja, jotka esiintyvät samoissa teksti-ikkunoissa. Näin muodostetut avainryppäät säilytetään. (Broglio, Callan, Croft, 1994.)

Kyselyn evaluointiprosessi asettaa dokumentit relevanssijärjestykseen käänteistiedostojen ja kyselyn perusteella. **Relevanssiarvioprosessi** muokkaa alkuperäistä kyselyä käyttäjän hakutulokselle antaman arvion perusteella. Kyselyn muokkaamiseen käytetään oppivia tekniikoita. Uusia avaimia ja fraaseja tunnistetaan relevanteista dokumenteista. Ne lisätään alkuperäiseen kyselyyn ja kaikki hakuavaimet painotetaan uudelleen. (Broglio, Callan, Croft, 1994.)

InQueryn arkkitehtuuri koostuu kolmesta alijärjestelmästä: jäsentäminen, tiedoston kääntäminen ja haku. Dokumenttiverkon rakentamisen ensimmäinen vaihe on kunkin dokumentin yhdistäminen oikeisiin sisällön esityssolmuihin. Tätä tehtävää kutsutaan **dokumentin jäsentämiseksi.** Toiminto koostuu viidestä osasta: sanaston analyysi, lauserakenteiden analyysi, käsitteiden tunnistaminen, sanakirjaan tallentaminen ja yhteyden muodostaminen, eli tunnistetun avaimen sijainnin tallentaminen. Dokumenttiverkko on siis jäsentämällä muodostunut yhteystiedostojen joukko. Jäsentämisen jälkeen verkko järjestetään ja siitä luodaan **käänteistiedosto**. (Callan, Croft, Harding, 1992.)

Hakualijärjestelmä kääntää kyselytekstin kyselyverkoksi ja evaluoi kyselyverkon aikaisemmin muodostetun dokumenttiverkon pohjalta. Käyttäjä voi antaa kyselyn joko luonnollisella kielellä tai rakenteisella kyselykielellä. Käyttäjän luonnollisella kielellä antamat kyselyt muokataan järjestelmälle käyttökelpoiseen muotoon käyttämällä sum-operaattoria:

$$\#sum(hakuavain_1, hakuavain_2, \ldots, hakuavain_n).$$

Kyselyavaimet muutetaan alkamaan pienellä alkukirjaimella. Sulkusanastoon kuuluvat sanat poistetaan. Avaimet stemmataan mahdollisimman yleiseen perusmuotoon ja verrataan käsitesanakirjaan ennen kyselyverkostoon siirtämistä. (Callan, Croft, Harding, 1992).

InQueryllä tapahtuvaa tiedonhakua on tutkittu paljon. Sen etuihin kuuluu laaja operaattorijoukko ja mahdollisuus painottaa hakuavaimia. Sen on todettu toimivan hyvin monissa kokeissa. (Kekäläinen, 1999.) Kyselyjen rakenteen on todettu olevan tärkeä tekijä, kun kyselyä laajennetaan. Paras tulos on saavutettu laajentamalla fasettimuotoisia kyselyjä (Järvelin & Kekäläinen 2000).

5.3 Synonyymien lähteet

Kokeissani laajensin kyselyt kahdella sanastolla: kaupallisella, yleisluontoisella Finthes-synonyymisanastolla sekä Kekäläisen (1999, 59) väitöskirjatutkimuksessaan TUTKiin räätälöimällä hierarkkisella käsitetesauruksella. Kutsun sitä tässä työssä tesaurukseksi. Tesaurus koostuu käsitteistä, ilmauksista ja täsmäytysmalleista ja niiden välisistä suhteita. Tesauruksen käsitteiden lukumäärä on 832 ja niiden ilmausten määrä on 1345. Täsmäytysmalleja eli ohjeita ilmaisun täsmäyttämiseen dokumenttiavaimiin on 1558. 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.)

5.4 Kyselyt

TUTK-tietokannassa on 35 aihetta, joihin on olemassa relevanssiarviot. Otin työhöni samat 30, joita Kekäläinen (1999) käytti. Vertasin viidenlaisia kyselyjä:

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

Kyselyt esitetään liitteessä 2. Kyselyjen rakennetta on selitetty sivulla 8.

Peruskyselyistä kahdelle (aiheet 6 ja 7) ei löytynyt lainkaan synonyymeja Finthesistä. Samoin tesauruksesta ei löytynyt laajennustermejä kahdelle peruskyselylle, aiheille 13 ja 22. Pidin nämä neljä aihetta silti mukana työssäni, sillä myös todellisissa hakutilanteissa on täysin mahdollista, ettei laajennusavaimia löydy.

Tein kyselyjä kaikista neljästä Sormusen hakuaiheluokasta. Kyselyt jakautuivat aiheittain seuraavasti:

- Aihe 8 kpl
- Rajattu aihe 9 kpl
- Henkilö 4 kpl
- Organisaatio 9 kpl

Henkilöaiheita on mukana selvästi muita tyyppejä vähemmän. Niistä suuri osa karsiutui pois jo Kekäläisen (1999) tutkimuksessa laajennettavuuden perusteella. Henkilöaiheiden laajentaminen on harvoin hyödyllistä; henkilön nimi on usein häntä koskevissa artikkeleissa niin hyvä hakuavain, että pelkästään nimellä saadaan erittäin hyvä hakutulos.

Peruskyselyjen pohjana käytin Kekäläisen (1999, 152) käsitteellisiä kyselysuunnitelmia. Kekäläinen ositti yhdyssanat omassa työssään, mutta oman työni yksinkertaistamiseksi päätin tässä työssä käsitellä yhdyssanoja kokonaisuutena. Tämä vaikutti myös laajentamiseen. Kekäläinen tutki myös yhdyssanojen osien laajentamista. Omissa laajennoksissani on mukana vain kokonaisten yhdyssanojen synonyymit.

5.4.1 Finthesillä laajentaminen

Mallinsin laajennosten tekemisessä automaattista laajennusta. 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 toisen 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.

Aiheiden 2 (velka) ja 23 (paast) kehitys-avaimelle Finthes-laajennoksiin synonyymeiksi tulivat mm. tulla, parannus ja aikuistua. Hakuavain kehitys sai Finthesistä seuraavat tulkinnat ja synonyymit:

kehitys (verbi)
muodostus, kehkeydys, muotoudus, sukeudus, synnys, tules
kehitys (verbi)
parannus, edistys, kohennus, kohenes
kasvas (verbi)
aikuistus, kehitys, kypsys, vartus
kasvu (subst.)
kehitys, kasvaminen, kehittyminen, kypsyminen
evoluutio (subst.)
kehitys
kehitys (subst.)
edistys, parannus, kohennus

Näistä Fintwol hyväksyi sellaisenaan tai muodosti perusmuodot:

muodostus kehkeytyä muotoutua sukeutua syntyä tulla parannus edistys kohennus koheta kasvaa aikuistua kypsyä varttua kasvaminen kehittyminen kypsyminen evoluutio

Näistä sanoista siis tuli kehitys-hakuavaimen laajennukset Finthes-kyselyihin.

5.4.2 Tesauruksella laajentaminen

Tesaurus on Kekäläisen (1999, 59) väitöskirjatutkimuksessaan 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

suojella vesi
vesi suojeleminen
vesistönsuojelu
suojella vesistö
vesistö suojeleminen.

5.4.3 Rakenteiset kyselyt

Rakenteiset kyselyt muodostin käyttämällä sum-operaattoria kokoavana operaattorina ja yhdistämällä synonyymifasetit syn-operaattorilla. Sumilla yhdistetyt avaimet ja fasetit ovat kyselyssä yhdenvertaisia. Sum-solmun arvo on avainten painojen keskiarvo. Syn taas käsittelee sisältämiään avaimia tai fasetteja saman avaimen esiintyminä. (Applied Computing Systems Institute of Massachusetts, 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.)

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

5.4.4 Sanaliittojen käsittely Finthes-laajennoksissa

Sanaliitot muodostin peruskyselyissä läheisyysoperaattorilla uwn, unordered window n. Operaattori edellyttää kaikkien hakuavaimien esiintyvän n:n sanan kokoisessa ikkunassa missä järjestyksessä tahansa (Applied Computing Systems Institute of Massachusetts, Inc.,1996). 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, jonka ikkunan kooksi valitsin operaattorin yhdistämien hakuavainfasettien lukumäärän lisättynä yhdellä ja pyöristettynä seuraavaan parittomaan lukuun. 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 kansanedustuslaitos eduskunta))…);
```

Päädyin tähän sanaliittojen käsittelymenetelmään tekemäni pikatestin perusteella. Viidessä kyselyssä, numeroissa 5 (varso), 11 (eyval), 14 (saksa) ja 20 (aids) esiintyy kussakin yksi sanaliitto, johon tuli laajennettava osuus. Kyselyssä 30 (vihr) on sanaliitto (vihreä liitto), jonka molemmat osat laajenivat. Kyselyssä 5 sanaliitto on (varsovan liitto), jossa synonyymeja tuli osalle liitto. Kyselyn 11 erisnimiliitto (ey parlamentti) sai synonyymeja osalle parlamentti. Kyselyssä 14 nimi Iso-Britannia sai käyttämäni sanaliittojen muodostusperiaatteen perusteella InQueryssä sanaliittomuodon (iso britannia). Synonyymeja löytyi osalle iso. Kyselyssä 20 esiintyy sanaliitto (ey maa), jossa synonyymeja kertyi sanalle maa.

Yksi vaihtoehto laajentaa sanaliitot on käyttää litteää synonyymifasettia, eli osat yhdistetään samanarvoisina syn-operaattorilla, esimerkiksi

#syn(iso runsas melkoinen aikamoinen hyvä reilu sievoinen suuri huomattava kova roima tuhti kookas isokokoinen järeä mittava suurikokoinen varteva aikuinen täysikasvuinen aikuisikäinen täysi-ikäinen britannia)

ja

#syn(vihreä kokematon tottumaton äkkinäinen outo aloitteleva puolue työryhmä joukkue ryhmä tiimi)

Vertasin tätä menetelmää siihen, että yhdistin sanaliiton osat uwn-operaattorilla, johon ikkunan kooksi otin fasettien lukumäärän lisättynä yhdellä, ja sidoin synonyymifasetin syn-operaattorilla:

#uw3(#syn(iso runsas melkoinen aikamoinen hyvä reilu sievoinen suuri huomattava kova roima tuhti kookas isokokoinen järeä mittava suurikokoinen varteva aikuinen täysikasvuinenaikuisikäinen täysi-ikäinen) britannia)

ja

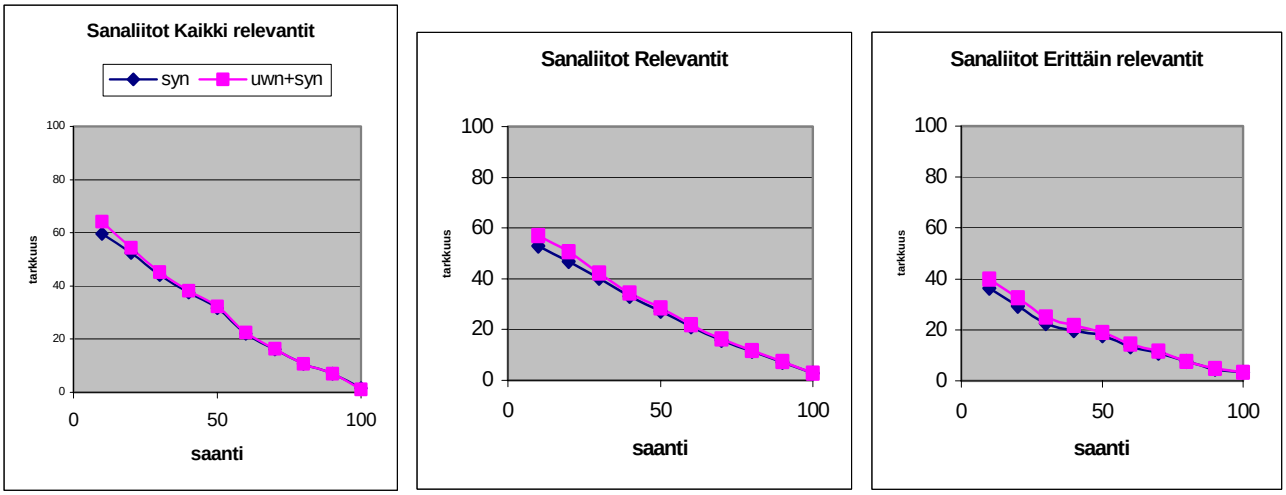
#uw3(#syn(vihreä kokematon tottumaton äkkinäinen outo aloitteleva) #syn(puolue työryhmä joukkue ryhmä tiimi))

Taulukossa 2 syn+uwn-menetelmän tarkkuuksien keskiarvojen keskiarvo on 23,99 ja syn-menetelmän 15,78. Syn+uwn-menetelmä tuottaa paremman tarkkuuden 11 tapauksessa ja syn-menetelmä neljässä. Parhaiten syn+uwn-menetelmästä hyötyy sanaliitto Varsovan Liitto, jolla laajennusmenetelmien välillä on eroa 23,31-42,19 prosenttiyksikköä eri relevanssitasoilla. Ainoa sanaliitto, jossa syn-laajennus toimii paremmin kaikilla relevanssitasoilla, on Iso-Britannia. Erot tosin ovat varsin pieniä, 0,12-0,26 prosenttiyksikköä.

Varsovan Liitto-sanaliiton menetelmien välinen ero kasvaa, kun relevanssitaso nousee. EY-parlamentilla menetelmien välinen ero pienenee relevanssitason noustessa. EY-parlamentin vaihtelu on suurin, 0,33-10,90 prosenttiyksikköä. EY-maan ja Vihreän Liiton muutokset eivät kehity johdonmukaisesti. Jos siis viiden sanaliiton tarkkuuksien keskiarvoja kolmella relevanssitasolla voi käyttää osviittana, syn+uwn-menetelmä näyttäisi löytävän sanaliitot tarkemmin.

Taulukko 2: Sanaliittojen käsittelyn vertailu: keskiarvotarkkuudet

kysely	rel.taso	syn	uwn+syn	erotus %-yks	sanaliitto
5	**Kaikki**	1,59	24,90	23,31	Varsovan liitto
	Relev.	1,85	37,36	35,51	
	Erittäin	2,83	45,02	42,19	
11	**Kaikki**	9,25	20,15	10,90	EY-parlamentti
	Relev.	5,73	11,96	6,23	
	Erittäin	5,62	5,95	0,33	
14	**Kaikki**	50,57	48,86	1,71	Iso-Britannia
	Relev.	45,74	44,16	1,58	
	Erittäin	30,29	30,02	0,27	
20	**Kaikki**	20,72	20,46	0,26	EY-maa
	Relev.	15,75	15,87	0,12	
	Erittäin	16,60	16,80	0,20	
30	**Kaikki**	12,23	13,83	1,60	Vihreä Liitto
	Relev.	15,19	15,97	0,78	
	Erittäin	2,67	8,58	5,91	
keskiarvo		15,78	23,99	8,21	
suurempia		4	11		



Kuvio 6: Sanaliiton käsittelyvaihtoehtojen saanti-tarkkuus -käyrät

Taulukko 3: Sanaliittojen käsittelyvaihtoehtojen vertailu

	Keskiarvotarkkuudet		Erotus	Tilastollinen merkitsevyys
taso	syn	uwn+syn	%-yks	P=
Kaikki	18,9	25,6	6,7	0,2188
Relevantit	16,9	25,1	8,2	0,0938
Erittäin	11,6	21,3	9,7	0,0938

Taulukko 3 kertoo, että sanaliittojen muotoilu syn+uwn-menetelmällä hyödyttää eniten erittäin relevanttien dokumenttien löytymistä. Sparck Jonesin peukalosännön mukaan kaikki erot kuuluvat keskiluokkaan. Kuvion 6 mukaan ero on suurin tulosjoukon alkupäässä, jossa käytännön merkityskin on suurin.

Sanaliittoja on vain viidessä kyselyssä kolmestakymmenestä. Tutkin sanaliittojen käsittelyn tilastollista merkitsevyyttä Wilcoxonin testillä (ks. Siegel 1989, 87). Kaikkiaan sanaliittokyselyjä on 15, viisi kyselyä kolmella relevanssitasolla. Kaikkien viidentoista kyselyn tuloseron merkitsevyystaso on 0,0177. Tätä eroa voidaan pitää melko merkitsevänä. Jos siis oletetaan, että syn+uwn -menetelmä on tehokkaampi kuin syn-menetelmä, väärässä olemisen todennäköisyys on 1,8 prosentin luokkaa. Relevanssitasoilla kaikkien relevantit virhetodennäköisyys on 22 prosenttia, ja korkeammilla tasoilla 9 prosenttia. Näistä mikään ei ole tilastollisesti huikean merkitsevä. Kun kyselyjä on vain 15, erojen tulisi olla todella suuria, jotta tilastollisia merkitsevyyksiä syntyisi.

5.5 Menetelmät

5.5.1 Saanti ja tarkkuus

Saanti ja tarkkuus ovat nykyään yleisimmät tiedonhaun tehokkuuden mittarit (Ks. esim. Salton & McGill, 1983; Järvelin 1995). Saanti kuvaa, kuinka paljon relevanteista dokumenteista löydettiin. Tarkkuus kuvaa, paljonko löydetyistä dokumenteista oli relevantteja. Saanti ja tarkkuus lasketaan seuraavasti:

$$saanti = \frac{löydettyjen\ relevanttien\ dokumenttien\ lukumäärä}{kokoelman\ relevanttien\ dokumenttien\ lukumäärä}$$

$$tarkkuus = \frac{löydettyjen\ relevanttien\ dokumenttien\ lukumäärä}{löydettyjen\ dokumenttien\ lukumäärä}$$

Tulos annetaan yleensä joko prosentteina tai lukuna väliltä [0,1]. Yleinen käytäntö on tarkastella tarkkuutta saannin funktiona kymmenellä saantitasolla taulukkona ja käyränä. Näin tehdään myös tässä työssä.

Tiedonhakijan tiedontarpeen luonne voi vaihdella. Joskus on tarpeen löytää mitä tahansa tiedonmurusia aiheesta, toisinaan taas tiedontarve on hyvin eksakti ja kaikki muu paitsi erittäin relevantti informaatio on hakijalle ajanhukkaa. Järvelin & Kekäläinen (2000) löysivät eroja hakumenetelmien välille erittäin relevanttien ja vain vähän relevanttien dokumenttien löytäjinä. Erittäin relevanttien dokumenttien löytäjinä kyselyrakenteiden tehokkuuden ero oli suurempi kuin vain vähän relevanttien dokumenttien. Hakijan kannalta olisi tärkeätä löytää hakumenetelmiä, jotka löytäisivät tulosjoukon alkupäähän relevanteimmat dokumentit. Tässä tutkimuksessa saantia ja tarkkuutta tarkastellaan erikseen kolmella relevanssitasolla: relevanteiksi katsotaan 1) kaikki relevantit, 2) melko ja erittäin relevantit tai 3) vain erittäin relevantit dokumentit (relevanssitasojen kuvaus s. 39).

5.5.2 Kumuloitu hyöty

Käyttäjän kannalta olisi mukavaa, 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. Käyttäjää hyödyttäisi eniten järjestelmä, joka näyttää tulosjoukon kärjessä kaikkein relevanteimmat dokumentit. Järjestelmän kykyä löytää erittäin relevantit dokumentit voidaan arvioida, kun dokumenttien relevanssi on arvioitu monitasoisesti ja eri relevanssitasojen hakutulosta verrataan keskenään.

Kaksi hyödyllistä mittaria, jotka mittaavat järjestelmän kykyä saada relevanteimmat dokumentit tulosjoukon kärkeen, ovat kumuloitu hyöty (cumulated gain, CG) ja alennettu kumuloitu

hyöty (discounted cumulated gain, DCG). (Järvelin & Kekäläinen, 2000a, Järvelin & Kekäläinen 2000b.) 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.

Esimerkissä relevanssiarvio on suoritettu neliportaisesti astein 0-3 (0 = ei-relevantti, 1 = vähän relevantti, 2 = melko relevantti, 3 = erittäin relevantti). Kun järjestysluvut korvataan relevanssiarvoilla, dokumenttivektori voi näyttää vaikka tältä:

G'=< 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, ... >

Kumuloitu hyöty paikassa i saadaan laskemalla yhteen arvot paikalla 1-i.

$$CG'=\begin{cases} G[1], jos\ i=1 \\ CG[i-1]+G[i] \end{cases}$$

Näin saadaan CG'=< 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, ...>. Siis 7. dokumentin kohdalla kumuloitu hyöty on 11. (Järvelin & Kekäläinen, 2000a, Järvelin & Kekäläinen 2000b.)

Mitä pidemmällä tulosjoukossa dokumentti tai viite sijaitsee, sitä vähemmän siitä on hyötyä hakijalle, koska hakija tuskin jaksaa selata muutamaa kymmentä dokumenttia enempää. Lisäksi hakijan tietämys kasvaa hänen hän tutustuessaan aihealueen kirjallisuuteen. Jos tämäkin halutaan ottaa huomioon tunnusluvussa, kumuloitua hyötyä voidaan pienentää sopivalla menetelmällä sitä enemmän, mitä suurempi alkuperäinen järjestysluku oli. Järvelin & Kekäläisen (2000a, 2000b) ehdottama menetelmä on jakaa kumuloitu hyöty sen järjestysluvun logaritmilla. Logaritmin kantaluvun valinnalla voidaan päättää, mallinnetaanko pikaista vai kärsivällistä hakijaa. Esimerkiksi logaritmin kantaluku 2 mallintaa kärsimätöntä tiedon tarvitsijaa, ja kantaluku 10 kärsivällistä.

$$DCG[i]=\begin{cases} G[i], jos\ i=1 \\ DCG[i-1]+G[i]/^{b}\log\ i \end{cases}$$

Tällöin DCG'=< 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, ... >

(Järvelin & Kekäläinen, 2000a, Järvelin & Kekäläinen 2000b.)

Hakujärjestelmän CG- ja DCG-arvoja voidaan verrata teoreettiseen parhaaseen mahdolliseen tulokseen. Paras mahdollinen vektori muodostetaan täyttämällä kunkin kyselyn vektorin alusta lähtien 3:lla niin monta paikkaa, kuin tämän aiheen kaikkein relevanteimpia dokumentteja on tietokannassa. Seuraavaksi vektoriin lisätään melko relevanttien dokumenttien lukumäärän verran arvoja 2. Lopuksi lisätään vähän relevanttien dokumenttien lukumäärän verran arvoja 1, ja loput paikat jätetään nolliksi. Tälle vektorille lasketaan CG ja DCG. (Järvelin & Kekäläinen, 2000a, Järvelin & Kekäläinen 2000b.)

Tässä työssä tarkastellaan sekä kumuloidun hyödyn että alennetun kumuloidun hyödyn tuloksia. Erittäin relevantit dokumentit painotetaan 10 kertaa arvokkaammiksi kuin vain vähän tai osittain relevantit dokumentit. Eri menetelmien kykyä muodostaa tällä lailla painotettu tulosjoukko verrataan.

5.5.3 Tilastolliset menetelmät

Tulosten tilastollista merkitsevyyttä tarkastelen Friedmanin kaksisuuntaisella järjestyslukutestillä (ks. esim. Siegel, 1989). Friedmanin testi on epäparametrinen 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 otosta. Otoksissa on oltava sama määrä tapauksia b. Muuttujan luokkien lukumäärä on k. Friedmanin testi selvittää, onko vertailtavilla otoksilla sama mediaani. Vertailtavista otoksista tehdään taulukko siten, että riville tulee tiedonhaun tutkimuksen tapauksessa yhden saantitason tarkkuudet eri menetelmillä. Saantiluvut muutetaan järjestysluvuiksi riveittäin. Testisuure lasketaan seuraavasti:

$$F_c = \frac{(b-1)(B_2 - bk(k+1)^2/4)}{A_2 - B_2}$$

missä

$$A_2 = \sum_{i=1}^{b} \sum_{j=1}^{k} (R(Xij))^2$$

ja

$$B_2 = \frac{1}{b}\sum_{j=1}^{k} Rj^2$$

b = rivien lukumäärä
k = sarakkeiden lukumäärä
$R(X_{ij})$ = solun järjestysluku i:nnellä rivillä j:nnessä sarakkeessa
R_j = j:nnen sarakkeen järjestyslukujen summa. (Conover 1980, tässä Kekäläinen 1999, 100.)

Kekäläinen (1999, 101) käytti tutkimuksessaan Conoverin versiota Friedmanin testistä, koska se osoittautui tehokkaammaksi kuin Siegelin (1989) versio. Omissa kokeissani käytän samoin Conoverin versiota. Conoverin versiossa testisuureen arvoa verrataan F-jakauman approksimaatiotaulukkoon. Testihypoteesi H_0 väittää, että kaikilla muuttujilla on sama mediaani. Jos H_0 voidaan hylätä Friedmanin testin perusteella, ainakin yksi menetelmä eroaa ainakin yhdesta muusta testatusta menetelmästä. Menetelmät i ja j ovat eri jakaumista, jos seuraava ehto täyttyy:

$$| R_j - R_i | \geq t_{1-\alpha/2}\left[\frac{2b(A_2 - B_2)}{(b-1)(k-1)}\right]^{1/2}$$

missä
R_j ja R_i ovat j:nnen ja i:nnen sarakkeen järjestyslukujen summat
A_2, B_2 ja k on annettu edellä
$t_{1-\alpha/2}$ on raja-arvo, jonka yläpuolella on $1-\alpha/2$ prosenttia jakaumasta.
(Conover 1980, 300; tässä Kekäläinen 1999, 101.)

6. Tulokset

Koetulosten tarkastelu on jaettu kahteen osaan. Ensimmäisessä osassa tarkastellaan kolmen eritasoisen relevanssikorpuksen hakutulosten saantia ja tarkkuutta. Friedmanin testillä arvioidaan tulosten tilastollista merkitsevyyttä ja Sparck Jonesin peukalosäännöllä tulosten käytännön merkitystä. Toisessa osassa hakutuloksia arvioidaan kumuloidun hyödyn menetelmillä.

6.1 Saanti ja tarkkuus ja tilastollinen merkitsevyys

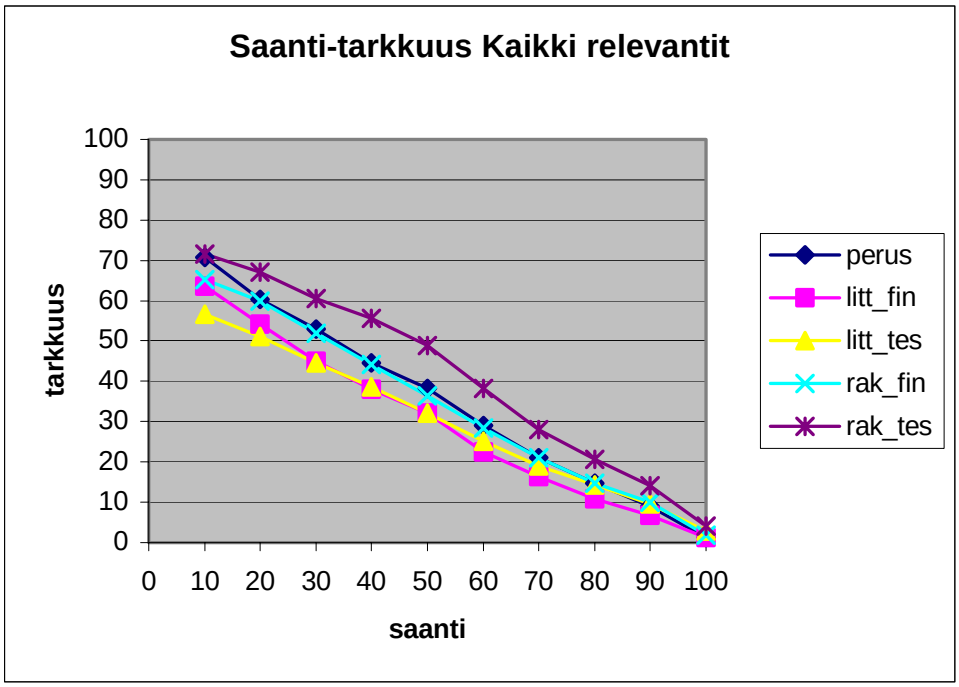
Työn tässä osassa menetelmien tehokkuutta arvioidaan saannin ja tarkkuuden avulla relevanssikorpuksittain. Myös tulosten tilastollista merkitsevyyttä ja käytännön merkitystä pohditaan tässä osassa.

6.1.1 Kaikki relevantit

Hakumenetelmien välille löytyy erittäin merkitseviä eroja tässä relevanssikorpuksessa. Jopa Friedmanin testi antoi tunnusluvuksi 0. Taulukosta 4 ja kuviosta 7 ilmenee, että kaikilla saantitasoilla rakenteisen tesauruslaajennuksen tarkkuus on paras. Huonoin tarkkuus on vaihtelee alhaisilla saantitasoilla litteiden laajennosten välillä. Korkeilla saantitasoilla ja 11 tason keskiarvotarkkuuden perusteella litteä Finthes-laajennos on heikoin kyselytyyppi.

Litteä laajentaminen on siis selvästi epäedullinen laajentamismenetelmä, ja rakenteinen tesauruslaajennos selvästi edullinen menetelmä. Rakenteisen tesauruslaajennoksen keskiarvotarkkuus on ainoa peruskyselyä parempi keskiarvotarkkuus. Yllättävää on, että TUTK-tesauruksella rakenteisesti laajentaminen on kokeen paras menetelmä, ja samalla sanastolla litteästi laajentaminen on toisiksi huonoin menetelmä.

Peruskyselyn ja rakenteisen tesauruskyselyn välinen ero on 6,7 prosenttiyksikköä, millä ei Sparck Jonesin mukaan ole käytännön merkitystä. Molempien litteiden menetelmien ja peruskyselyn välinen ero jää myös keskiluokkaan. Litteät menetelmät eivät siis peukalosäännön mukaan ole selkeästi huonompia kuin peruskysely, mutta ero on kuitenkin kiinnostava. Rakenteisen Finthes-laajennuksen ja litteiden laajennusten väliset erot jäävät alle viiden prosenttiyksikön, joten niillä ei



Kuvio 7: Kaikki relevantit saanti—tarkkuus -käyrä

Taulukko 4: Saanti—tarkkuus: kaikki relevantit

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 5: Friedmanin testi: kaikki relevantit

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

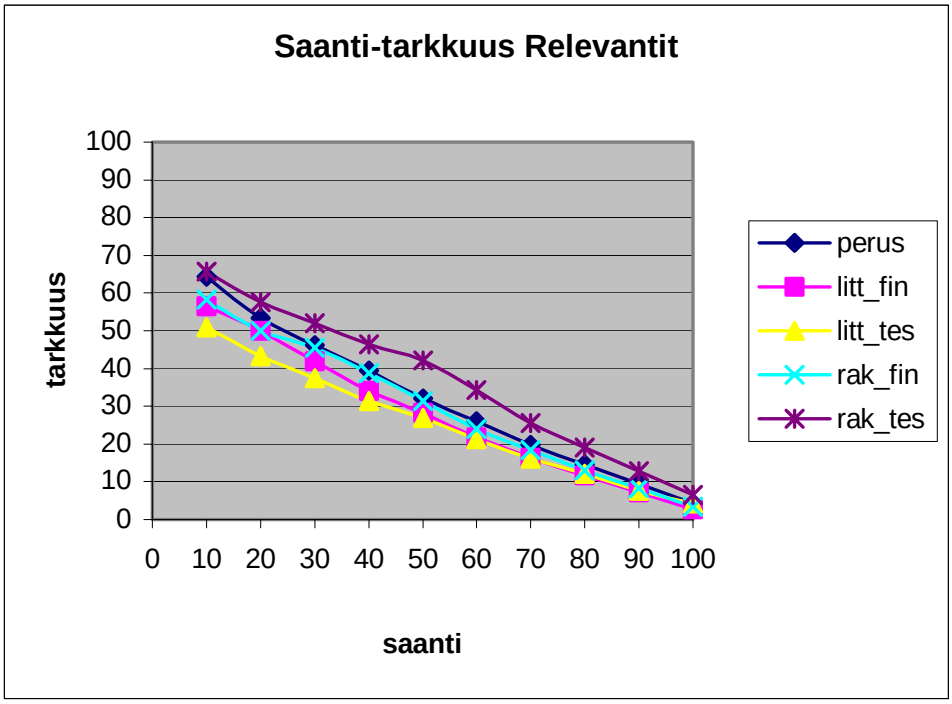
Taulukossa 5
- = ei merkitsevää eroa
* = p<0,05 melko merkitsevä ero
** = p<0,005 varsin merkitsevä ero
*** = p<0,001 erittäin merkitsevä ero

ole käytännön eroa. Peukalosäännön perusteella ainoa käytännön merkitystä osoittava ero on litteiden laajennusten ja rakenteisen tesauruslaajennuksen välillä.

Friedmanin testin mukaan peruskysely on litteää Finthes-kyselyä tilastollisesti erittäin merkitsevästi (5,2 prosenttiyksikköä) parempi. Menetelmien välisten erojen tilastollinen merkitsevyys näkyy taulukossa 5. Suurimmat tilastolliset erot löytyvät litteiden ja rakenteisten kyselyjen välillä, rakenteisten eduksi. Peruskyselyn ja rakenteisen Finthes-laajennuksen välillä ei ole merkitsevää tilastollista eroa, mutta rakenteinen tesauruslaajennus on perusmenetelmää varsin merkitsevästi parempi. Rakenteinen Finthes-laajennus on melko merkitsevästi huonompi menetelmä kuin rakenteinen tesauruslaajennus. Sparck Jonesin peukalosäännöllä rakenteisen Finthes-laajennuksen ja molempien litteiden laajennusmenetelmien välinen ero ei ollut edes kiinnostava, mutta Friedmanin testin mukaan ero on tilastollisesti erittäin merkitsevä.

6.1.2 Relevantit dokumentit

Relevanttien dokumenttien Friedmanin testin tunnusluku on 0,000000002 eli menetelmien väliset erot ovat erittäin merkittäviä. Tälläkin tasolla rakenteinen tesauruslaajennus on systemaattisesti



Kuvio 8: Relevantit saanti—tarkkuus –käyrä

kaikilla saantitasoilla paras menetelmä (ks. kuvio 8 ja taulukko 6). Huonoin menetelmä on 10-70 % saannilla litteä tesauruslaajennus ja korkeammilla tasoilla litteä Finthes-laajennus. Litteän tesauruslaajennuksen keskiarvotarkkuus on huonoin. Ainoa peruskyselyä parempi keskiarvotarkkuus on rakenteisella tesauruslaajennuksella.

Sparck Jonesin peukalotuntumalla käytännössä tärkeä ero on vain litteän ja rakenteisen tesauruslaajennuksen välillä. Kaikki muut erot ovat käytännössä vain joko kiinnostusta herättäviä tai merkityksettömiä. Rakenteisten menetelmien keskinäinen ero on kiinnostava, litteiden menetelmien keskinäinen ero merkityksetön. Peruskyselyyn nähden ero molempiin Finthes-laajennoksiin on merkityksetön (molemmat huonompia) ja molempiin tesauruslaajennoksiin kiinnostava (litteä laajennus on huonompi ja rakenteinen parempi menetelmä).

Friedmanin testi osoitti, että erot kahdella ylemmällä relevanssitasolla eivät ole ihan yhtä merkitseviä kuin kaikkien relevanttien korpuksessa. Suurimmat p-arvot ovat litteiden menetelmien ja rakenteisen tesauruslaajennuksen välillä. Litteiden laajennusten keskinäinen ero ei ole tilastollisesti merkitsevä, mutta rakenteisten menetelmien välinen ero on varsin merkitsevä.

Taulukko 6: Saanti—tarkkuus: relevantit

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

Rakenteisista menetelmistä Finthes-laajennus ei ole merkitsevästi peruskyselyä huonompi, mutta tesauruslaajennus on sitä varsin merkitsevästi parempi. Molemmat litteät laajennusmenetelmät on merkitsevästi peruskyselyä huonompia.

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.

Taulukko 7: Friedmanin testi: kaikki relevantit

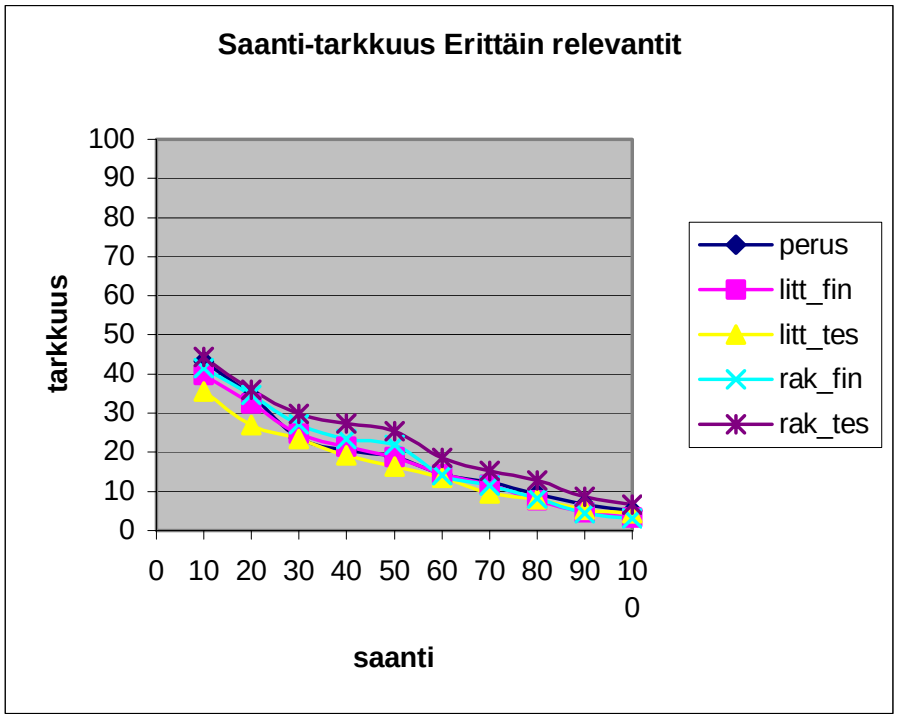
	perus	litt_fin	litt_tes	rak_fin
litt_fin	**			
litt_tes	**	-		
rak_fin	-	**	**	
rak_tes	*	***	***	**

Taulukossa 6
- = ei merkitsevää eroa
* = p<0,05 melko merkitsevä ero
** = p<0,005 varsin merkitsevä ero
*** = p<0,001 erittäin merkitsevä ero

6.1.3 Erittäin relevantit dokumentit

Erittäin relevanttien dokumenttien korpuksessa menetelmien välillä on entistä vähemmän eroa. Saantikantojen koko tällä tasolla on selvästi pienempi kuin muissa korpuksissa (ks. liite 3), joten yhden relevantin dokumentin löytyminen tai löytymättä jääminen on suhteessa tärkeämpää kuin suuremmissa relevanssikorpuksissa. Friedmanin testin tunnusluku oli 0,0000469 eli tälläkin tasolla on silti erittäin merkitseviä eroja.

Tälläkin tasolla rakenteinen tesauruslaajennus on paras hakumenetelmä kaikilla saantitasoilla (taulukko 7, kuvio 9). Samoin kuin tasolla relevantit, litteä tesauruslaajennus on huonoin menetelmä 70 prosentin saantitasolle asti ja litteä Finthes-laajennos 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. Tällä tasolla keskiarvotarkkuuksien perusteella molemmat rakenteiset kyselytyypit toimivat paremmin kuin peruskysely, joskaan ero peruskyselyn ja rakenteisen Finthes-laajennuksen välillä ei ole tilastollisesti merkitsevä (taulukko 8) eikä systemaattinen (taulukko 7 ja kuvio 9) eikä Sparck



Kuvio 9: Erittäin relevantit saanti—tarkkuus –käyrä

Taulukko 8: Saanti-tarkkuus: erittäin relevantit

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

Jonesin peukalotuntumalla edes kiinnostava. Yli 10 prosenttiyksikön eroja tällä menetelmällä ei syntynyt yhtään ja 5-10 prosenttiyksikön eroja vain litteän ja rakenteisen tesauruslaajennuksen välille. Mikään menetelmä ei ole edes kiinnostusta herättävästi parempi tai huonompi kuin peruskysely. Tilastollisesti tälläkään relevanssitasolla litteiden laajennusten välinen ero ei ole merkitsevä, mutta rakenteisten menetelmien välinen ero on varsin merkitsevä. Peruskysely on rakenteista tesauruslaajennusta melko merkitsevästi huonompi. Molempien litteiden menetelmien huonommuus rakenteiseen tesauruslaajennukseen nähden on erittäin merkitsevä.

Taulukko 9: Friedmanin testi: erittäin relevantit

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

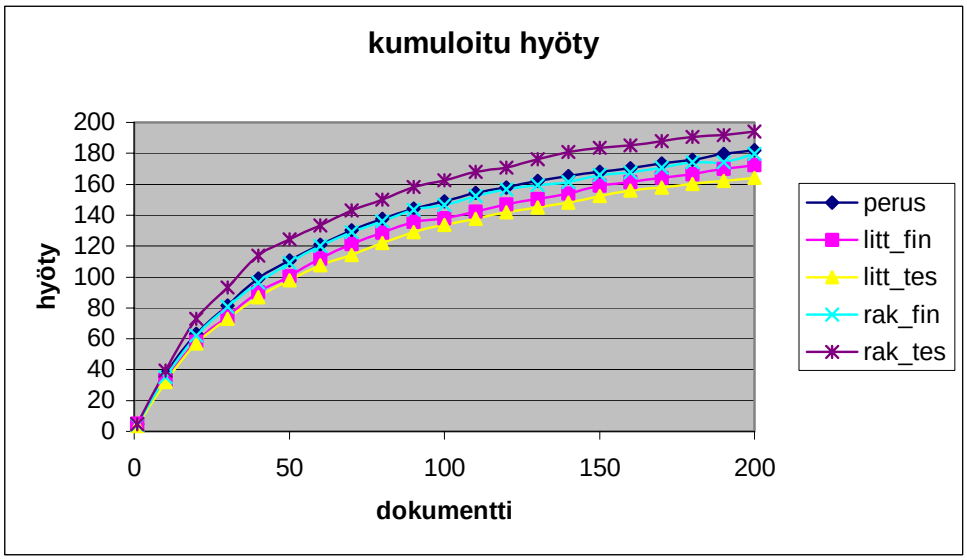
Taulukossa 9
- = ei merkitsevää eroa
* = p<0,05 melko merkitsevä ero
** = p<0,005 varsin merktisevä ero
*** = p<0,001 erittäin merkitsevä ero

6.2 Kumuloidun hyödyn menetelmät

6.2.1 Kumuloitu hyöty

Käytin kumuloidun hyödyn laskemisessa erittäin relevanteille dokumenteille painoa 1, relevanteille painoa 5 ja melko relevanteille painona 10. Erittäin relevantti dokumentti oli siis kymmenen kertaa arvokkaampi kuin melko relevantti dokumentti. Logaritmin kantalukuna oli 2 eli mallinsin kärsimätöntä käyttäjää (Järvelin & Kekäläinen, 2000), joita suurin osa tavallisista tiedonhakijoista todennäköisesti on. 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 10). Huonoin menetelmä läpi koko listan on litteä tesauruslaajennus. Kuviosta 10 käy ilmi, että rakenteisen tesauruslaajennuksen jälkeen paras menetelmä on peruskysely, mutta ihan sen kyljessä kulkee litteä Finthes-laajennus.

Kumuloidusta hyödystä voidaan päätellä myös se, että sama hyöty, kuin mikä peruskyselyllä saavutetaan 20 dokumenttia selaamalla, saadaan rakenteisella tesauruskyselyllä jo 17. dokumentin kohdalla. Litteällä tesauruskyselyllä sama hyöty edellyttää 24 dokumentin tutkimisen. Löytääkseen saman määrän relevantteja dokumentteja kuin peruskyselyn 20 dokumentilla, tiedonhakijan



Kuvio 10: Kumuloitu hyöty

Taulukko 10: Kumuloitu hyöty

	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

tarvitsee selata vain 17 dokumenttia, jos kysely on muodostettu rakenteisesti tesauruksesta laajentamalla. Jos kysely on laajennettu samasta tesauruksesta litteästi, tiedon hakija joutuu selaamaan 24 dokumenttia löytääkseen yhtä paljon relevantteja.

Taulukko 11: Kumuloitu hyöty: erot prosenttiyksikköinä

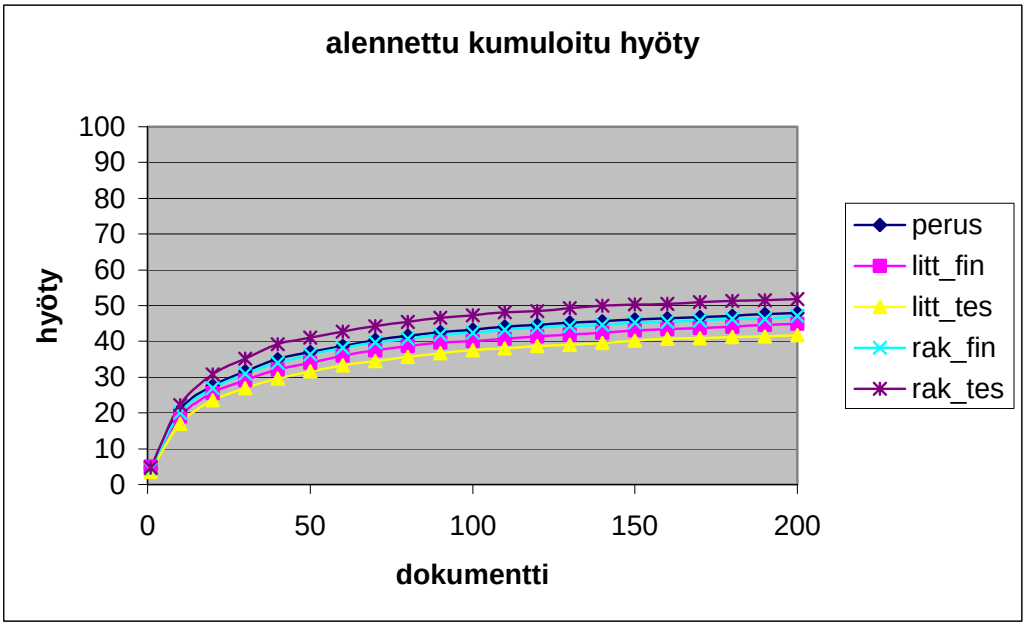
	perus	litt_fin	litt_tes	rak_fin
litt_fin	8,8			
litt_tes	13,6	4,8		
rak_fin	2,0	6,8	11,6	
rak_tes	12,6	21,0	25,8	14,2

Sparck Jonesin peukalosäännön perusteella litteä tesauruslaajennus on peruskyselyä käytännössäkin selvästi huonompi menetelmä (ks taulukko 11). Käytännössä ainoa peruskyselyä parempi menetelmä on rakenteinen tesauruslaajennus. Peruskyselyn ja litteän Finthes-kyselyn välinen ero on kiinnostava. Rakenteinen tesauruslaajennus on kaikkia muita menetelmiä parempi myös käytännössä. Lisäksi rakenteinen Finthes-laajennus on käytännössäkin selvästi parempi menetelmä kuin litteä tesauruslaajennus.

6.2.2 Alennettu kumuloitu hyöty

Menetelmien väliset erot ovat pienemmät alennettu kumuloitu hyöty -menetelmällä laskettuna, kuin kumuloitu hyöty -menetelmällä. Käyrät kulkevat kiinni toisissaan kuviossa 11.

Taulukosta 12 käy ilmi, että ensimmäisen dokumentin paras menetelmä on litteä Finthes-laajennus, ja sen jälkeen jälleen rakenteinen tesauruslaajennus on paras menetelmä loppuun saakka, ja sama sanasto litteästi laajennettuna huonoin menetelmä. Peruskysely on jälleen parempi vaihtoehto kuin mikään muu menetelmä paitsi rakenteinen tesauruslaajennos. Alennettu kumuloitu hyöty kertoo, että peruskyselyn 20. dokumentin hyöty saavutetaan rakenteisella tesauruskyselyllä 16. dokumentin kohdalla ja litteällä tesauruskyselyllä 32. dokumentin kohdalla.



Kuvio 11: Alennettu kumuloitu hyöty

Taulukko 12: Alennettu kumuloitu hyöty

	perus	litt_fin	litt_tes	rak_fin	rak_tes
1	4,6	5,0	3,5	4,3	4,8
10	20,9	19,1	17,0	20,0	22,2
20	27,6	25,5	23,5	26,9	30,9
30	31,6	29,0	26,9	30,9	35,2
40	34,9	32,1	29,6	33,9	39,2
50	37,1	34,0	31,6	36,3	41,1
60	38,8	36,0	33,3	38,2	42,7
70	40,4	37,5	34,4	39,7	44,3
80	41,6	38,7	35,7	40,8	45,4
90	42,6	39,8	36,7	41,9	46,7
100	43,3	40,1	37,5	42,4	47,3
110	44,1	40,8	38,1	43,3	48,1
120	44,6	41,5	38,7	43,9	48,5
130	45,2	42,0	39,1	44,3	49,3
140	45,7	42,5	39,6	44,6	50,0
150	46,1	43,2	40,2	45,2	50,4
160	46,4	43,5	40,7	45,4	50,6
170	46,8	43,8	40,9	45,8	51,0
180	47,2	44,2	41,2	46,4	51,3
190	47,7	44,6	41,5	46,4	51,5
200	48,0	44,9	41,8	47,1	51,8
Keskiarvo	39,3	36,6	33,9	38,5	43,0

Taulukko 13: Alennettu kumuloitu hyöty: erot prosenttiyksikköinä

	perus	litt_fin	litt_tes	rak_fin
litt_fin	2,7			
litt_tes	5,4	2,7		
rak_fin	0,8	1,9	4,6	
rak_tes	3,7	6,4	9,1	4,5

Taulukko 13 kertoo, että alennettu kumuloitu hyöty ei löydä käytännössä tärkeitä eroja menetelmien välillä Sparck Jonesin peukalotuntumalla. Peruskysely on parempi kuin litteä tesauruslaajennus, mutta ero on vain kiinnostava. Rakenteinen tesauruslaajennus on samassa luokassa molempia litteitä menetelmiä parempi. Kaikki muut erot jäävät alle huomiorajan. Peruskyselyn ja muiden menetelmien välillä ei siis ole käytännössä eroa alennetun kumuloidun hyödyn perusteella.

7. Keskustelu ja johtopäätökset

7.1 Keskustelua tuloksista

Kristensenin (1992) tutkimuksessaan käyttämä hakujärjestelmä perustui Boolen logiikkaan, joten hänen tuloksensa eivät ole suoraan vertailukelpoisia omieni kanssa. Tutkimuksessa rinnakkaistermilaajennus oli hyödyllisin litteiden kyselyjen laajentamismenetelmä.

Voorheesin (1994) litteiden kyselyjen tutkimuksessa vain lyhyiden kyselyjen tulos parani merkitsevästi laajentamalla. Oma kokeiluni on sikäli vertailukelpoinen Voorheesin tulosten kanssa, että Voorheesin järjestelmä oli vektoripohjainen eli osittaistäsmäyttävä, kuten käyttämäni todennäköisyyslaskentaan perustuva InQuerykin. 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. 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.

Liitteessä 4 näkyy kunkin aiheen keskiarvotarkkuudet eri menetelmillä. Taulukoista huomaa, että kukin testattu menetelmä käy niin ykkös- kuin häntäpaikalla. Rakenteisen tesauruslaajennuksen parhaimmuus on suurin kaikki relevantit -tasolla. Relevanssitason noustessa menetelmien väliset erot tasoittuvat. Tämä tulos näkyy myös Friedmanin testin tuloksista. Friedmanin testi lasketaan tarkkuuksien keskinäisten järjestysten perusteella. Järvelinin ja Kekäläisen (2000) tutkimuksessa verrattaessa erilaisia kyselyn rakenteita ja painotusmenetelmiä erot kasvavat siirryttäessä matalalta relevanssitasolta korkeammalle. Omassa tutkimuksessani kävi päinvastoin: erot ovat selvästi suurempia kaikkien relevanttien korpuksessa kuin erittäin relevanttien dokumenttien. Tämä kertoo, että synonyymilaajennus ei välttämättä auta saamaan erittäin relevantteja dokumentteja tulosjoukon kärkeen. Tätä päätelmää tukevat myös kumuloitu hyöty- ja alennettu kumuloitu hyöty -menetelmien tulokset. Alennettu kumuloitu hyöty antaa suuremman arvon tulosjoukon alkupäässä oleville

dokumenteille. Tällä menetelmällä menetelmien väliset erot ovat pienempiä kuin kumuloidulla hyödyllä mitattaessa.

Yleisellä tasolla tuloksista on helppo vetää ainakin se nopea johtopäätös, että 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 liene kovin vaikeaa. Toinen yhtä itsestäänselvä tulos oli, että rakenteisesti laajennettaessa tietokantaa varten räätälöity sanasto on ehdottomasti paras laajennusavainten lähde. Yleisellä synonyymisanastolla 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.

Tuloksissa esiintyy mielenkiintoinen johdonmukaisuus: rakenteinen tesauruslaajennus on kaikissa testeissä kokeen paras menetelmä ja litteä tesauruslaajennus on kokeen huonoin menetelmä kaikilla muilla mittareilla paitsi kaikki relevantit -korpuksen keskiarvotarkkuudella. Tosin litteän tesaurus- ja litteän Finthes-laajennuksen välinen ero ei ole millään relevanssitasolla tilastollisesti merkitsevä eikä käytännössä mielenkiintoinen..

Kyselyjä, joissa litteä Finthes-laajennus päihitti litteän tesauruslaajennuksen oli yhteensä 9. Ne jakautuivat Sormusen hakuaiheluokkiin taulukon 14 kertomalla tavalla. Eniten niitä kyselyjä, joissa litteä Finthes-laajennus toimii paremmin, on rajattu aihe -luokassa (5/9 eli 55,6 %). Vähiten voittoisa litteä Finthes-laajennus on organisaatio-luokassa (2/9 eli 22,2 %). Henkilö-luokassa näitä tapauksia on 1/4 eli 25 % ja aihe-luokassa 3/8 eli 37,5 %. Ero voi johtua siitä, että rajattu aihe -luokassa hakuaihe on mahdollista rajata kohtuullisen tarkoin hakuavaimilla, ja laajentamisen tulosta on vaikea ennustaa. Toisaalta litteiden laajennusmenetelmien välinen ero ei ole millään relevanssitasolla tilastollisesti merkitsevä, joten ero voi johtua myös sattumasta.

Tarkastelin myös Kekäläisen (1999, 152) määrittelemien pää- ja sivukäsitteiden lukumääriä ja niiden saamia synonyymeja. Tuloksia näkyy taulukossa 15. Alkuperäisessä kyselyssä on 71 pääkäsitettä edustavaa avainta ja 72 sivukäsitettä edustavaa avainta. Finthesillä laajentamalla

Taulukko 14: Aiheluokittain litteä Finthes parempi kuin litteä tesaurus

Aiheluokka	Kyselyjä luokassa kaikkiaan	Kysely
Aihe	8	**3**
Rajattu aihe	9	**4**
		6
		7
		10
		20
Henkilö	4	**1**
Organisaatio	9	**5**
		16

pääkäsiteavainten määrä nousee 143:een ja tesauruksella laajentamalla 208:aan. Sivukäsiteavainten määrä taas nousee Finthesillä 281:een ja tesauruksella 166:een. Finthesillä pääkäsiteavaimille löytyy 2,01 synonyymia ja sivukäsiteavaimille 3,9. Tesaurus toimii päinvastaisella tavalla, pääkäsiteavaimille löytyy enemmän synonyymeja (2,93) kuin sivukäsiteavaimille (2,31). Ero ei tesauruksella ole yhtä suuri kuin Finthesillä. Tämä ero voisi selittää sen, miksi tesaurus toimii paremmin rakenteisesti laajentamalla. Kekäläinen on kokeellisesti määritellyt pääkäsitteet ratkaisevan tärkeiksi hakutuloksen kannalta. Pääkäsiteavaimille annetut hyvät synonyymit parantavat hakutulosta enemmän kuin sivukäsiteavainten synonyymit.

Taulukko 15: Hakuavainten ja synonyymien määrät

	perus	finthes	tesaurus
Pääkäsitettä edustavia avaimia	71	143	208
Sivukäsitettä edustavia avaimia	72	281	166
Pääkäsiteavaimia / kysely	2,37	4,77	6,93
Sivukäsiteavaimia / kysely	2,40	9,37	5,53
Pääkäsiteavaimille synonyymeja / hakuavain		2,01	2,93
Sivukäsiteavaimille synonyymeja / hakuavain		3,90	2,31

7.2 Johtopäätökset

Aluksi tekemäni sanaliittokokeen perusteella sanaliitot kannattaa muodostaa syn+uwn-menetelmällä. 15 kyselyllä menetelmien välinen ero osoittautui tilastollisesti melko merkitseväksi ja käytännössä kiinnostavaksi.

Kaikilla menetelmillä tuli varsin selväksi, että tutkituista vaihtoehdoista vain rakenteisesti laajentaminen on jonkin verran hyödyllistä. Kahdella relevanssitasolla kolmesta ja molemmilla kumuloitua hyötyä mittaavilla menetelmillä peruskysely on rakenteisen tesauruslaajennuksen jälkeen toisiksi paras menetelmä. Peruskyselyn ja rakenteisen Finthes-laajennuksen välinen ero ei ole tilastollisesti merkitsevä millään menetelmällä, mutta peruskysely saa silti parempia tuloksia kuin rakenteinen Finthes-laajennus kaikilla muilla menetelmillä mitattuna kuin erittäin relevantit -korpuksen keskiarvotarkkuudella.

Voorheesin, Kekäläisen ja omien kokeideni tulosten perusteella voi päätellä, että pelkät synonyymit eivät ole riittävä laajennusavainlähde ainakaan litteästi laajentamalla. Rakenteisestikin laajentamisesta on hyötyä vain, jos synonyymit on peräisin kokoelmaa varten räätälöidystä sanastosta. Rakenteinen tesauruslaajennus on Sparck Jonesin peukalosäännön mukaan peruskyselyä vain kiinnostavasti parempi kahdella ensimmäisellä relevanssitasolla, vain erittäin relevantit dokumentit sisältävässä korpuksessa ero on merkityksetön. Tilastollisesti rakenteinen tesauruslaajennus on peruskyselyä varsin merkitsevästi parempi kaikkien relevanttien dokumenttien korpuksessa ja vain melko merkitsevästi parempi relevanttien ja erittäin relevanttien dokumenttien tasolla. Kumuloitua hyötyä rakenteinen tesauruslaajennus tuottaa 12,6 prosenttiyksikköä enemmän kuin peruskysely, mikä on peukalosäännön mukaan jo käytännössä merkittävä ero, mutta alennettu kumuloitu hyöty osoittaa todellisen käytännön hyödyn olevan vain 3,7 prosenttiyksikköä eli merkityksetön.

Mitään syytä laajentaa litteästi tai Finthesillä tämä työ ei siis löydä ja rakenteisen tesauruslaajennuksenkin ero peruskyselyyn on niin vähäinen, että todelliseksi tiedonhakujärjestelmän parantajaksi siitä tuskin on. Jatkossa olisi mielenkiintoista tutkia kyselyn rakenteen vaikutusta kieltenvälisessä tiedonhaussa tai hakujärjestelmän käyttöliittymän vaikutusta vuorovaikutteisen kyselynlaajentamisen tulokseen.

8. Lähteet

Kirjallisuus

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

Alkula, R. (2000). Merkkijonoista suomen kielen sanoiksi. Suomenkielisten morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja liittämisen vaikutukset tekstin tallennukseen ja hakuun. Acta Universitatis Tamperensis 763. Tampere: University of Tampere. Saatavana myös www-muodossa: <http://acta.uta.fi/pdf/951-44-4886-3.pdf> Käytetty 18.2.2002.

Belkin, N. J. & Croft, W. B. 1987. Retrieval Techniques. Annual Review of Information Science and Technology, vol. 22. Elsevier Science Publishers B. V., 109–145.

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.

Cosijn, E., Ingwersen, P. (2000). Dimensions of relevance. Information Processing and Management. 36(2000) 533-550.

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

Greenberg, J. (2001). Automatic Query Expansion via Lexical-Semantic Relationships. Journal of the American Society for Information Science and Technology 52(5), 402-415.

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

Järvelin, K., Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM, 41 - 48.

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>.

Karlsson, F. (1998). Yleinen kielitiede. Helsinki: Yliopistopaino. Helsingin yliopisto.

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.

Kekäläinen, J., Järvelin, K. (2000). The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. Information Retrieval 2000(1): 329-344. Saatavilla myös www-muodossa: <http://www.info.uta.fi/tutkimus/fire/archive/JK&KJ-IR'00.pdf> Käytetty 22.3.2002.

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

Korfhage, R. R. (1997). Information storage and retrieval. John Wiley & Sons, Inc.

Mandala, R., Tokunaga, T., Tanaka, H. (2000). Query expansion using heterogeneous thesauri. Information Processing and Management 36(2000), 361-378.

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.

Nie J-Y, Simard M, Isabelle P, Durand R. (1999). Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM, 74-81

Pirkola, A. (2001). Morphological typology of languages for IR. Journal of Documentation, 57(3), 330-348.

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.

Sarasevic, T. (1996). Relevance Reconsidered. Information Science: Integration in perspectives. Proceedings of the Second Conference on Conceptions of Librar and Information Science (CoLIS 2). Copenhagen (Denmark). 201-218. Saatavana myös www-muodossa osoitteessa: <http://www.scils.rutgers.edu/~tefko/CoLIS2_1996.doc>. Käytetty 1.7.02

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).

Swanson, D. (1988). Historical Note: Information Retrieval and the Future of an Illusion. Journal of the American Society for Information Science 39(4), 92-98.

Tietohuollon sanasto: suomi. ruotsi, englanti, saksa, ranska. (1993). Tekniikan sanastokeskus ja Tietopalveluseura. Helsinki: Kirjastopalvelu 1993.

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.

Wilson, P. (1977). Public knowledge, private ignorance. Contributions in librarianship and information science; no 10. Westport, Connecticut: Greenwood Press, Inc.

Verkkolähteet

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

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>

Porter, M. The Porter Stemming Algorithm. Käytetty 3.10.2002.
IRL:<http://www.tartarus.org/~martin/PorterStemmer/>

Porter, M. Snowball, Käytetty 3.10.2002. IRL:< http://snowball.tartarus.org/>

Muut lähteet

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

LIITTEET

Liite 1: Hakuaiheet

numero	lyhenne	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 ympäristöystävällisempien 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 ympäristönsuojelukysymyksenä. 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.

Liite 2: Kysely

2.1 Peruskyselyt

```
#q1 = #sum(#uw3(george bush) #uw3(mihail gorbatshov)
helsinki tapaaminen sopimus päätös);

#q2 = #sum(#uw3(etelä amerikka) velkakriisi velkaantumisongelma
kehitys ratkaisu);

#q3 = #sum(metsäteollisuus polkumyynti yhdysvallat
oikeudenkäynti));

#q4 = #sum(jyväskylä kaupunki maalaiskunta kuntaliitos
talous kannattaja vastustaja);

#q5 = #sum(#uw3(varsova liitto) lopettaminen jäsenmaa päätös);

#q6 = #sum(liettua taloussaarto neuvostoliitto lopettaminen);

#q7 = #sum(irak joukkotuhoase yk hävittäminen inventoiminen);

#q8 = #sum(opec öljy tuotanto hinta päätös);

#q9 = #sum(bukarest kaivosmies väkivalta oppositio);

#q10 = #sum(namibia rauhanturvaoperaatio @untag yk itsenäistyminen);

#q11 = #sum(#uw3(ey parlamentti) päätöksenteko #uw3(ey
toimielin));

#q12 = #sum(#uw3(carl @bildt) pohjoismaa yhteistyö lausunto);

#q13 = #sum(jugoslavia presidenttineuvosto istunto päätös);

#q14 = #sum(saksa yhdistyminen miehittäjävalta yhdysvallat
#uw3(iso britannia) ranska neuvostoliitto neuvottelu sopimus);

#q15 = #sum(valmet metsäkone maansiirtokone traktori
kiskokalusto tuotanto kannattavuus);

#q16 = #sum(tampella irtisanominen työvoima);

#q17 = #sum(kera @ktm matkailu laina avustus investointi);

#q18 = #sum(neste maakaasu hankinta jakelu markkinointi);

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

#q20 = #sum(#uw3(ey maa) aids levinneisyys);

#q21 = #sum(elintarviketeollisuus tuontirajoitus poisto
suomi);

#q22 = #sum(asuntotuotanto suhdanne ennuste tilasto);

#q23 = #sum(liikenne päästö katalysaattori lainsäädäntö
kehitys);

#q24 = #sum(#uw3(japani autoteollisuus)
#uw3(eurooppalainen autonvalmistaja) autotehdas investointi
 yhteistyö);
```

```
#q25 = #sum(#uw3(kemiallinen metsäteollisuus) vesiensuojelu
investointi);

#q26 = #sum(kauppa aukioloaika järjestö ammattijärjestö
pidentäminen mielipide);

#q27 = #sum(pakkaus kierrätys lainsäädäntö);

#q28 = #sum(#uw3(esko aho) #uw3(ey jäsenyyshakemus)
mielipide);

#q29 = #sum(#uw3(kauko juhantalo) ydinvoima ratkaisu
mielipide);

#q30 = #sum(#uw3(vihreä puolue) eduskunta kansanedustaja
aloite välikysymys puheenvuoro äänestys);
```

2.2 Litteät Finthes-kyselyt

```
#q1 = #sum(#uw3(george bush) #uw3(mihail gorbatshov) helsinki
tapaaminen sopimus sitoumus kontrahti sopimuskirja päätös valinta ratkaisu);

#q2 = #sum(#uw3(etelä amerikka) velkakriisi velkaantumisongelma
kehitys muodostus kehkeytyä muotoutua sukeutua syntyä tulla
parannus edistys kohennus koheta kasvaa aikuistua kypsyä
varttua kasvaminen kehittyminen kypsyminen evoluutio
ratkaisu valinta päätös);

#q3 = #sum(metsäteollisuus polkumyynti dumping dumppaus
yhdysvallat oikeudenkäynti prosessi oikeudenistunto);

#q4 = #sum(jyväskylä kaupunki maalaiskunta kunta pitäjä
kuntaliitos talous kotitalous huusholli rahatalous
taloudenhoito ekonomia varainhoito kannattaja tukija
puoltaja liittolainen vastustaja vihollinen vihamies
oppositio vastustuspuolue kilpailija kilpakumppani
antagonisti vastavaikuttaja);

#q5 = #sum(#uw3(varsova #syn(liitto organisaatio järjestö
liittoutuma liittouma yhteenliittymä)) lopettaminen
jäsenmaa päätös valinta ratkaisu);

#q6 = #sum(liettua taloussaarto neuvostoliitto lopettaminen);

#q7 = #sum(irak joukkotuhoase yk hävittäminen inventoiminen);

#q8 = #sum(opec öljy tuotanto teollisuus produktio hinta raha
käteinen fyrkka fyffe hynä mani nappula rahna saldo tuohi
valuutta kurssi arvo taksa päätös valinta ratkaisu);

#q9 = #sum(bukarest kaivosmies väkivalta pahoinpitely oppositio
vastustus vastarinta vastustaja vastustuspuolue vastakkaisuus
vastakohta);

#q10 = #sum(namibia rauhanturvaoperaatio @untag yk itsenäistyminen
emansipaatio vapautuminen emansipoituminen);

#q11 = #sum(#uw3(ey #syn(parlamentti kansanedustuslaitos eduskunta))
päätöksenteko #uw3(ey toimielin));

#q12 = #sum(#uw3(carl @bildt) pohjoismaa yhteistyö yhteistoiminta
koordinaatio yhteispeli lausunto);

#q13 = #sum(jugoslavia presidenttineuvosto istunto päätös valinta
ratkaisu);

#q14 = #sum(saksa yhdistyminen miehittäjävalta yhdysvallat
#uw3(#syn(iso runsas melkoinen aikamoinen hyvä reilu sievoinen suuri
huomattava kova roima tuhti kookas isokokoinen järeä mittava
suurikokoinen varteva aikuinen täysikasvuinen aikuisikäinen täysi-ikäinen)
britannia) ranska neuvostoliitto neuvottelu konsultaatio neuvonkysyntä
kokous palaveri keskustelu väittely sopimus sitoumus kontrahti sopimuskirja);

#q15 = #sum(valmet metsäkone maansiirtokone traktori kiskokalusto tuotanto
teollisuus produktio kannattavuus);

#q16 = #sum(tampella irtisanominen työvoima henkilökunta henkilöstö);

#q17 = #sum(kera kanssa mukana myötä ohella seura @ktm matkailu turismi
matkailuharrastus laina velka luotto vippi avustus apuraha stipendi
```

```
tuki tukiainen investointi sijoitus);

#q18 = #sum(neste maakaasu luonnonkaasu hankinta ostos jakelu levitys
jako markkinointi myynti tarjonta);

#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);

#q20 = #sum(#uw3(ey #syn(maa valtio valtakunta valta multa maaseutu bönde
lande provinssi maaperä maapohja maankamara maapallo tellus maanpinta
kamara tantere tanner)) aids immuunikato levinneisyys);

#q21 = #sum(elintarviketeollisuus tuontirajoitus poisto suomi piiskata
 ruoskia vitsoa);

#q22 = #sum(asuntotuotanto suhdanne konjunktuuri ennuste prognoosi
tilasto statistiikka);

#q23 = #sum(liikenne trafiikki päästö katalysaattori lainsäädäntö
legislaatio oikeusjärjestys kehitys muodostus kehkeytyä muotoutua
sukeutua syntyä tulla parannus edistys kohennus koheta kasvaa
aikuistua kypsyä varttua kasvaminen kehittyminen kypsyminen evoluutio);

#q24 = #sum(#uw3(japani autoteollisuus) #uw3(eurooppalainen autonvalmistaja)
autotehdas investointi sijoitus yhteistyö yhteistoiminta koordinaatio
yhteispeli);

#q25 = #sum(#uw3(kemiallinen metsäteollisuus) vesiensuojelu investointi
sijoitus);

#q26 = #sum(kauppa myymälä boutique liike puoti putiikki shop liiketoiminta
kaupankäynti bisnes aukioloaika kauppajärjestö ammattijärjestö
pidentäminen mielipide kanta ajatus kannanotto käsitys näkemys);

#q27 = #sum(pakkaus rasia lipas aski kotelo laatikko käärö paketti kierrätys
uudelleenkäyttö jälleenkäyttö lainsäädäntö legislaatio oikeusjärjestys);


#q28 = #sum(#uw3(esko #syn(aho niitty keto)) #uw3(ey jäsenyyshakemus)
mielipide kanta ajatus kannanotto käsitys näkemys);

#q29 = #sum(#uw3(kauko juhantalo) ydinvoima ydinenergia ratkaisu valinta
päätös mielipide kanta ajatus kannanotto käsitys näkemys);

#q30 = #sum(#uw3(#syn(vihreä kokematon tottumaton äkkinäinen outo aloitteleva)
#syn(puolue työryhmä joukkue ryhmä tiimi)) eduskunta kansanedustuslaitos
parlamentti kansanedustaja aloite ehdotus esitys vireillepano välikysymys
puheenvuoro suunvuoro äänestys vaali);
```

2.3 Litteät tesauruskyselyt

```
#q1 = #sum(#uw3(george bush) #uw3(presidentti bush) bush
#uw3(mihail gorbatshov) #uw3(presidentti gorbatshov) gorbatshov
helsinki tapaaminen tavata sopimus päätös);

#q2 = #sum(#uw3(etelä amerikka) #uw3(latinalainen amerikka)
@lattarimaat @lattarimaiden velkakriisi velkaantumisongelma
kehitys kehittyä kehkeytyminen kehkeytyä kehittyminen ratkaisu);

#q3 = #sum(metsäteollisuus puunjalostusteollisuus
 puuteollisuus polkumyynti dumping dumpingmyynti
dumppaus hinnanpolkeminen #uw3(hinta polkeminen)
yhdysvallat usa amerikka oikeudenkäynti oikeusjuttu
oikeuskäsittely käräjät);

#q4 = #sum(jyväskylä kaupunki maalaiskunta kuntaliitos
#uw3(kunta liittäminen) #uw3(liittää kunta)
#uw3(kunta yhdistäminen) #uw3(yhdistää kunta)
talous kannattaja puoltaja vastustaja);

#q5 = #sum(#uw3(varsova liitto) #uw3(varsova sotilasliitto)
lopettaminen lopetus lopettaa lakkauttaminen lakkautus
lakkauttaa jäsenmaa jäsenvaltio päätös);

#q6 = #sum(liettua taloussaarto taloussulku talousboikotti
#uw3(saartaa talous) neuvostoliitto nl lopettaminen lopetus
lopettaa lakkauttaminen lakkautus lakkauttaa);

#q7 = #sum(irak joukkotuhoase massatuhoase yk #uw3(yhdistynyt
kansakunta) hävittäminen hävitys hävittää eliminoida
tuhoaminen tuhota  #uw3(raivata pois) inventoiminen
inventointi inventoida);

#q8 = #sum(opec #uw3(öljyntuottajamaa järjestö)
#uw3(öljynviejämaa järjestö) öljy tuotanto hinta päätös);

#q9 = #sum(bukarest kaivosmies kaivostyöläinen
väkivalta oppositio #uw3(hallitus vastustaja)
 #uw3(vastustaa hallitus));

#q10 = #sum(namibia #uw3(lounais afrikka)
rauhanturvaoperaatio @untag @untag @untagia @untagiin
@untagilla @untagille @untagilta @untagin @untagissa
@untagjoukkojen @untagkin) yk #uw3(yhdistynyt kansakunta)
 itsenäistyminen itsenäistyä);

#q11 = #sum(#uw3(ey parlamentti) #uw5(eurooppa yhteisö parlamentti)
#uw3(yhteisö parlamentti) #uw5(eurooppa unioni parlamentti)
 päätöksenteko #uw3(tehdä päätös) #uw3(ey toimielin)
#uw5(eurooppa yhteisö toimielin) #uw3(yhteisö toimielin)
#uw5(eurooppa unioni toimielin));

#q12 = #sum(#uw3(carl @bildt) #uw3(carl @bildtiä)
#uw3(carl @bildtiin) #uw3(carl @bildtillä)
#uw3(carl @bildtille) #uw3(carl @bildtiltä) #uw3(carl @bildtin)
#uw3(carl @bildtistä)
#uw3(puheenjohtaja @bildt) #uw3(puheenjohtaja @bildtiä)
#uw3(puheenjohtaja @bildtiin) #uw3(puheenjohtaja @bildtillä)
#uw3(puheenjohtaja @bildtille) #uw3(puheenjohtaja @bildtiltä)
#uw3(puheenjohtaja @bildtin) #uw3(puheenjohtaja @bildtistä)
#uw3(pääministeri @bildt) #uw3(pääministeri @bildtiä)
#uw3(pääministeri @bildtiin) #uw3(pääministeri @bildtillä)
```

```
#uw3(pää ministeri @bildtille) #uw3(pääministeri @bildtiltä)
#uw3(pääministeri @bildtin) #uw3(pääministeri @bildtistä)
@bildt @bildtiä @bildtiin @bildtillä @bildtille @bildtiltä
@bildtin @bildtistä pohjoismaa skandinavia yhteistyö
yhteistoiminta lausunto);

#q13 = #sum(jugoslavia presidenttineuvosto istunto päätös);

#q14 = #sum(saksa yhdistyminen yhdistyä liittyminen
liittyä miehittäjävalta yhdysvallat usa #uw3(iso britannia)
englanti britannia ranska neuvostoliitto nl neuvottelu
neuvotella neuvotteleminen sopimus);

#q15 = #sum(valmet metsäkone maansiirtokone traktori
kiskokalusto ratakalusto raidekalusto tuotanto kannattavuus);

#q16 = #sum(tampella irtisanominen irtisanoa #uw3(sanoa
irti) #uw3(vähentää työvoima) #uw3(työvoima vähentäminen)
#uw3(vähentää työpaikka) #uw3(työpaikka vähentäminen))
työvoima);

#q17 = #sum(kera kehitysaluerahasto @ktm #uw5(kauppa ja
teollisuusministeriö) teollisuusministeriö matkailu turismi
 laina luotto avustus tuki tukiainen tukipalkkio tukiraha
subventio investointi investoiminen investoida sijoitus);

#q18 = #sum(neste maakaasu hankinta hankkiminen hankkia
jakelu markkinointi markkinoiminen markkinoida);

#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);

#q20 = #sum(#uw3(ey maa) #uw3(ey jäsen) #uw3(ey valtio)
#uw5(eurooppa yhteisö jäsen) #uw5(eurooppa yhteisö maa)
#uw5(eurooppa yhteisö valtio) #uw5(eurooppa yhteisö jäsenmaa)
#uw5(eurooppa yhteisö jäsenvaltio) aids immuunikato
#uw3(hiv tartunta) #uw3(hiv infektio) levinneisyys yleisyys);

#q21 = #sum(elintarviketeollisuus ruokateollisuus
tuontirajoitus tuontikiintiö tuontisuoja
tuontisäännöstely poisto poistaminen suomi);

#q22 = #sum(asuntotuotanto suhdanne ennuste tilasto);

#q23 = #sum(liikenne päästö katalysaattori lainsäädäntö
kehitys kehittyä kehkeytyminen kehkeytyä kehittyminen);

#q24 = #sum(#uw3(japani autoteollisuus)
#uw3(japanilainen autoteollisuus)
#uw3(eurooppalainen autonvalmistaja) autotehdas investointi
investoiminen investoida sijoitus yhteistyö yhteistoiminta);

#q25 = #sum(#uw3(kemiallinen metsäteollisuus)
#uw3(kemiallinen puunjalostusteollisuus) vesiensuojelu
#uw3(suojella vesi) #uw3(vesi suojeleminen) vesistönsuojelu
#uw3(suojella vesistö) #uw3(vesistö suojeleminen) investointi investoiminen
investoida sijoitus);

#q26 = #sum(kauppa kauppaliike aukioloaika aukiolo kauppajärjestö ammattijärjestö
pidentäminen pidentää jatkaa mielipide näkökanta näkökohta);
```

```
#q27 = #sum(pakkaus kääre päällinen kierrätys
kierrättäminen kierrättää lainsäädäntö);

#q28 = #sum(#uw3(esko aho) aho #uw3(pääministeri aho)
#uw3(puheenjohtaja aho) #uw3(ey jäsenyyshakemus)
#uw3(ey jäsenyysanomus) mielipide näkökanta näkökohta);

#q29 = #sum(#uw3(kauko juhantalo) juhantalo #uw3(ministeri
juhantalo)) ydinvoima ydinenergia atomivoima
atomienergia ratkaisu mielipide näkökanta näkökohta);

#q30 = #sum(#uw3(vihreä puolue) #uw3(vihreä liitto) eduskunta
kansanedustaja parlamentaarikko aloite ehdotus välikysymys
puheenvuoro äänestys);
```

2.4 Rakenteiset Finthes-kyselyt

```
#q1 = #sum(#uw3(george bush) #uw3(mihail gorbatshov) helsinki tapaaminen
#syn(sopimus sitoumus kontrahti sopimuskirja)
#syn(päätös valinta ratkaisu));

#q2 = #sum(#uw3(etelä amerikka) velkakriisi velkaantumisongelma
#syn(kehitys muodostus kehkeytyä muotoutua sukeutua syntyä tulla
parannus edistys kohennus koheta kasvaa aikuistua kypsyä varttua
kasvaminen kehittyminen kypsyminen evoluutio)
#syn(ratkaisu valinta päätös));

#q3 = #sum(metsäteollisuus #syn(polkumyynti dumping dumppaus) yhdysvallat
#syn(oikeudenkäynti prosessi oikeudenistunto));

#q4 = #sum(jyväskylä kaupunki #syn(maalaiskunta kunta pitäjä) kuntaliitos)
#syn(talous kotitalous huusholli rahatalous taloudenhoito ekonomia varainhoito)
#syn(kannattaja tukija puoltaja liittolainen) #syn(vastustaja vihollinen
vihamies oppositio vastustuspuolue kilpailija kilpakumppani antagonisti
vastavaikuttaja));

#q5 = #sum(#uw3(varsova #syn(liitto organisaatio järjestö liittoutuma liittouma
yhteenliittymä)) lopettaminen jäsenmaa #syn(päätös valinta ratkaisu));

#q6 = #sum(liettua taloussaarto neuvostoliitto lopettaminen);

#q7 = #sum(irak joukkotuhoase yk hävittäminen inventoiminen);

#q8 = #sum(opec öljy #syn(tuotanto teollisuus produktio)
#syn(hinta raha käteinen fyrkka fyffe hynä mani nappula rahna
saldo tuohi valuutta kurssi arvo taksa) #syn(päätös valinta ratkaisu));

#q9 = #sum(bukarest kaivosmies #syn(väkivalta pahoinpitely)
#syn(oppositio vastustus vastarinta vastustaja vastustuspuolue
vastakkaisuus vastakohta));

#q10 = #sum(namibia rauhanturvaoperaatio @untag yk
#syn(itsenäistyminen emansipaatio vapautuminen emansipoituminen));

#q11 = #sum(#uw3(ey #syn(parlamentti kansanedustuslaitos eduskunta))
päätöksenteko #uw3(ey toimielin));

#q12 = #sum(#uw3(carl @bildt) pohjoismaa
#syn(yhteistyö yhteistoiminta koordinaatio yhteispeli) lausunto);

#q13 = #sum(jugoslavia presidenttineuvosto istunto
#syn(päätös valinta ratkaisu));

#q14 = #sum(saksa yhdistyminen miehittäjävalta yhdysvallat
#uw3(#syn(iso runsas melkoinen aikamoinen hyvä reilu sievoinen suuri
huomattava kova roima tuhti kookas isokokoinen järeä mittava suurikokoinen
varteva aikuinen täysikasvuinen aikuisikäinen täysi-ikäinen) britannia)
ranska neuvostoliitto #syn(neuvottelu konsultaatio neuvonkysyntä
kokous palaveri keskustelu väittely)
#syn(sopimus sitoumus kontrahti sopimuskirja));

#q15 = #sum(valmet metsäkone maansiirtokone traktori kiskokalusto
#syn(tuotanto teollisuus produktio) kannattavuus);

#q16 = #sum(tampella irtisanominen #syn(työvoima henkilökunta henkilöstö));

#q17 = #sum(#syn(kera kanssa mukana myötä ohella seura) @ktm
#syn(matkailu turismi matkailuharrastus) syn(laina velka luotto vippi)
#syn(avustus apuraha stipendi tuki tukiainen) #syn(investointi sijoitus));
```

```
#q18 = #sum(neste #syn(maakaasu luonnonkaasu) #syn(hankinta ostos)
#syn(jakelu levitys jako) #syn(markkinointi myynti tarjonta));

#q19 = #sum(ydinvoimala ydinjäte #syn(käsittely työstö muokkaus
työstäminen manipulointi manipulaatio 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));

#q20 = #sum(#uw3(ey #syn(maa valtio valtakunta valta multa maaseutu bönde
lande provinssi maaperä maapohja maankamara maapallo tellus maanpinta
kamara tantere tanner)) #syn(aids immuunikato) levinneisyys);

#q21 = #sum(elintarviketeollisuus tuontirajoitus poisto
#syn(suomi piiskata ruoskia vitsoa));

#q22 = #sum(asuntotuotanto #syn(suhdanne konjunktuuri) #syn(ennuste prognoosi)
#syn(tilasto statistiikka));

#q23 = #sum(#syn(liikenne trafiikki) päästö katalysaattori
#syn(lainsäädäntö legislaatio oikeusjärjestys)
#syn(kehitys muodostus kehkeytyä muotoutua sukeutua syntyä tulla parannus
edistys kohennus koheta kasvaa aikuistua kypsyä varttua kasvaminen
kehittyminen kypsyminen evoluutio));

#q24 = #sum(#uw3(japani autoteollisuus) #uw3(eurooppalainen autonvalmistaja)
autotehdas #syn(investointi sijoitus)
#syn(yhteistyö yhteistoiminta koordinaatio yhteispeli));

#q25 = #sum(#uw3(kemiallinen metsäteollisuus) vesiensuojelu
#syn(investointi sijoitus));

#q26 = #sum(#syn(kauppa myymälä boutique liike puoti putiikki shop
liiketoiminta kaupankäynti bisnes) aukioloaika kauppajärjestö ammattijärjestö
pidentäminen #syn(mielipide kanta ajatus kannanotto käsitys näkemys));

#q27 = #sum(#syn(pakkaus rasia lipas aski kotelo laatikko käärö paketti)
#syn(kierrätys uudelleenkäyttö jälleenkäyttö)
#syn(lainsäädäntö legislaatio oikeusjärjestys));

#q28 = #sum(#uw3(esko #syn(aho niitty keto))
#uw3(ey jäsenyyshakemus)
#syn(mielipide kanta ajatus kannanotto käsitys näkemys));

#q29 = #sum(#uw3(kauko juhantalo) #syn(ydinvoima ydinenergia)
#syn(ratkaisu valinta päätös)
#syn(mielipide kanta ajatus kannanotto käsitys näkemys));

#q30 = #sum(#uw3(#syn(vihreä kokematon tottumaton äkkinäinen outo aloitteleva)
#syn(puolue työryhmä joukkue ryhmä tiimi))
#syn(eduskunta kansanedustuslaitos parlamentti) kansanedustaja
#syn(aloite ehdotus esitys vireillepano) välikysymys
#syn(puheenvuoro suunvuoro) #syn(äänestys vaali));
```

2.5 Rakenteiset tesauruskyselyt

```
#q1 = #sum(#syn(#uw3(george bush) #uw3(presidentti bush) bush)
#syn(#uw3(mihail gorbatshov) #uw3(presidentti
gorbatshov) gorbatshov) helsinki #syn(tapaaminen tavata) sopimus päätös);

#q2 = #sum(#syn(#uw3(etelä amerikka) #uw3(latinalainen amerikka)
@lattarimaat @lattarimaiden) velkakriisi velkaantumisongelma
#syn(kehitys kehittyä kehkeytyminen kehkeytyä kehittyminen)
ratkaisu);

#q3 = #sum(#syn(metsäteollisuus puunjalostusteollisuus
 puuteollisuus) #syn(polkumyynti dumping dumpingmyynti
dumppaus hinnanpolkeminen #uw3(hinta polkeminen))
#syn(yhdysvallat usa amerikka) #syn(oikeudenkäynti
oikeusjuttu oikeuskäsittely käräjät));

#q4 = #sum(jyväskylä kaupunki maalaiskunta #syn(kuntaliitos
#uw3(kunta liittäminen) #uw3(liittää kunta) #uw3(kunta
yhdistäminen) #uw3(yhdistää kunta))
talous #syn(kannattaja puoltaja) vastustaja);

#q5 = #sum(#syn(#uw3(varsova liitto) #uw3(varsova
sotilasliitto)) #syn(lopettaminen lopetus lopettaa
lakkauttaminen lakkautus lakkauttaa) #syn(jäsenmaa
jäsenvaltio) päätös);

#q6 = #sum(liettua #syn(taloussaarto taloussulku
talousboikotti #uw3(saartaa talous)) #syn(neuvostoliitto nl)
#syn(lopettaminen lopetus lopettaa lakkauttaminen lakkautus
lakkauttaa));

#q7 = #sum(irak #syn(joukkotuhoase massatuhoase) #syn(yk
#uw3(yhdistynyt kansakunta)) #syn(hävittäminen hävitys
hävittää eliminoida tuhoaminen tuhota #uw3(raivata pois))
#syn(inventoiminen inventointi inventoida));

#q8 = #sum(#syn(opec #uw3(öljyntuottajamaa järjestö)
#uw3(öljynviejämaa järjestö)) öljy tuotanto hinta päätös);

#q9 = #sum(bukarest #syn(kaivosmies kaivostyöläinen)
väkivalta #syn(oppositio #uw3(hallitus vastustaja)
#uw3(vastustaa hallitus)));

#q10 = #sum(#syn(namibia #uw3(lounais afrikka))
rauhanturvaoperaatio #syn(@untag @untag
@untagia @untagiin @untagilla @untagille @untagilta
@untagin @untagissa @untagjoukkojen @untagkin) #syn(yk
#uw3(yhdistynyt kansakunta)) #syn(itsenäistyminen
itsenäistyä));

#q11 = #sum(#syn(#uw3(ey parlamentti) #5(eurooppa yhteisö
parlamentti) #uw3(yhteisö parlamentti) #5(eurooppa unioni
parlamentti)) #syn(päätöksenteko #uw3(tehdä päätös))
#syn(#uw3(ey toimielin) #uw5(eurooppa yhteisö toimielin)
#uw3(yhteisö toimielin) #uw5(eurooppa unioni toimielin)));

#q12 = #sum(#syn(#uw3(carl @bildt) #uw3(carl @bildtiä)
#uw3(carl @bildtiin) #uw3(carl @bildtillä)
#uw3(carl @bildtille) #uw3(carl @bildtiltä) #uw3(carl @bildtin)
#uw3(carl @bildtistä)
#uw3(puheenjohtaja @bildt) #uw3(puheenjohtaja @bildtiä)
#uw3(puheenjohtaja @bildtiin) #uw3(puheenjohtaja @bildtillä)
#uw3(puheenjohtaja @bildtille) #uw3(puheenjohtaja @bildtiltä)
```

```
#uw3(puheenjohtaja @bildtin) #uw3(puheenjohtaja @bildtistä)
#uw3(pääministeri @bildt) #uw3(pääministeri @bildtiä)
#uw3(pääministeri @bildtiin) #uw3(pääministeri @bildtillä)
#uw3(pääministeri @bildtille) #uw3(pääministeri @bildtiltä)
#uw3(pääministeri @bildtin) #uw3(pääministeri @bildtistä)
@bildt @bildtiä @bildtiin @bildtillä @bildtille @bildtiltä
@bildtin @bildtistä) #syn(pohjoismaa skandinavia)
#syn(yhteistyö yhteistoiminta) lausunto);

#q13 = #sum(jugoslavia presidenttineuvosto istunto päätös);

#q14 = #sum(saksa #syn(yhdistyminen yhdistyä liittyminen
liittyä) miehittäjävalta #syn(yhdysvallat usa)
#syn(#uw3(iso britannia) englanti britannia) ranska
#syn(neuvostoliitto nl) #syn(neuvottelu neuvotella
neuvotteleminen) sopimus);

#q15 = #sum(valmet metsäkone maansiirtokone traktori
#syn(kiskokalusto ratakalusto raidekalusto) tuotanto kannattavuus);

#q16 = #sum(tampella #syn(irtisanominen irtisanoa #uw3(sanoa
irti) #uw3(vähentää työvoima) #uw3(työvoima vähentäminen)
#uw3(vähentää työpaikka) #uw3(työpaikka vähentäminen)) työvoima);

#q17 = #sum(#syn(kera kehitysaluerahasto) #syn(@ktm
#uw3(kauppa ja teollisuusministeriö) teollisuusministeriö)
 #syn(matkailu turismi) #syn(laina luotto avustus tuki
tukiainen  tukipalkkio tukiraha subventio) #syn(investointi
investoiminen investoida sijoitus));

#q18 = #sum(neste maakaasu #syn(hankinta hankkiminen
hankkia) jakelu #syn(markkinointi markkinoiminen markkinoida));

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

#q20 = #sum(#syn(#uw3(ey maa) #uw3(ey jäsen) #uw3(ey valtio)
#uw5(eurooppa yhteisö jäsen) #uw5(eurooppa yhteisö maa)
#uw5(eurooppa yhteisö valtio) #uw5(eurooppa yhteisö jäsenmaa)
#uw5(eurooppa yhteisö jäsenvaltio)) #syn(aids immuunikato
#uw3(hiv tartunta) #uw3(hiv infektio)) #uwsyn(levinneisyys
yleisyys));

#q21 = #sum(#syn(elintarviketeollisuus ruokateollisuus)
#syn(tuontirajoitus tuontikiintiö tuontisuoja
tuontisäännöstely) #syn(poisto poistaminen) suomi);

#q22 = #sum(asuntotuotanto suhdanne ennuste tilasto);

#q23 = #sum(liikenne päästö katalysaattori lainsäädäntö
#syn(kehitys kehittyä kehkeytyminen kehkeytyä kehittyminen));

#q24 = #sum(#syn(#uw3(japani autoteollisuus) #uw3(japanilainen
autoteollisuus)) #uw3(eurooppalainen autonvalmistaja)
autotehdas #syn(investointi investoiminen
investoida sijoitus) #syn(yhteistyö yhteistoiminta));

#q25 = #sum(#syn(#uw3(kemiallinen metsäteollisuus)
#uw3(kemiallinen puunjalostusteollisuus))
```

```
#syn(vesiensuojelu #uw3(suojella vesi)
#uw3(vesi suojeleminen) vesistönsuojelu #uw3(suojella
vesistö) #uw3(vesistö suojeleminen))
#syn(investointi investoiminen investoida sijoitus));

#q26 = #sum(#syn(kauppa kauppaliike) #syn(aukioloaika
aukiolo) kauppajärjestö ammattijärjestö
#syn(pidentäminen pidentää jatkaa)
#syn(mielipide näkökanta näkökohta));

#q27 = #sum(#syn(pakkaus kääre päällinen) #syn(kierrätys
kierrättäminen kierrättää) lainsäädäntö);

#q28 = #sum(#syn(#uw3(esko aho) aho #uw3(pääministeri aho)
#uw3(puheenjohtaja aho)) #syn(#uw3(ey jäsenyyshakemus)
#uw3(ey jäsenyysanomus))
#syn(mielipide näkökanta näkökohta));

#q29 = #sum(#syn(#uw3(kauko juhantalo) juhantalo #uw3(ministeri
juhantalo)) #syn(ydinvoima ydinenergia atomivoima
atomienergia) ratkaisu #syn(mielipide näkökanta näkökohta));

#q30 = #sum(#syn(#uw3(vihreä puolue) #uw3(vihreä liitto)) eduskunta
#syn(kansanedustaja parlamentaarikko)
#syn(aloite ehdotus) välikysymys puheenvuoro äänestys);
```

Liite 3: Relevanttien dokumenttien lukumäärä

Aihe	Kaikki rel.	Relevantit	Erittäin rel.
1	54	32	14
2	81	55	11
3	21	16	10
4	16	8	7
5	129	48	17
6	145	87	15
7	83	65	39
8	72	29	6
9	37	24	7
10	143	98	47
11	54	29	13
12	19	13	1
13	111	36	19
14	84	54	17
15	68	45	6
16	26	14	1
17	33	17	2
18	65	37	10
19	68	34	26
20	43	21	8
21	90	14	3
22	122	36	3
23	136	87	17
24	24	16	5
25	45	25	13
26	35	27	13
27	73	59	26
28	35	21	6
29	14	6	2
30	27	13	2
Yhteensä	**1953**	**1066**	**366**
keskiarvo	**65,1**	**35,5**	**12,2**

Liite 4: Keskiarvotarkkuudet tyypeittäin ja aiheittain

4.1 Kaikki relevantit

tyyppi	numero	aihe	perus	litt_fin	litt_tes	rak_fin	rak_tes
aihe	3	polku	21,6	46,3	32,5	51,2	52,4
	19	yjate	46,7	25,5	45,9	47,2	58,3
	21	elint	21,9	19,2	20,9	21,9	28,8
	22	asunt	8,2	8,0	8,2	8,3	8,2
	23	paast	42,4	28,1	30,2	44,7	42,7
	25	sellu	0,7	0,3	0,5	0,4	0,6
	26	aukio	12,8	11,2	25,2	12,1	30,6
	27	kierr	55,9	43,7	53,1	52,7	57,7
raj.aihe	2	velka	9,2	1,3	13,2	4,5	17,9
	4	jykul	51,3	42,1	35,4	0,4	45,0
	6	liett	73,9	73,9	34,8	73,9	73,7
	7	iraki	55,6	55,6	52,7	55,6	62,4
	9	bukar	54,7	46,5	62,3	55,7	72,8
	10	untag	83,2	82,6	34,2	83,6	85,1
	14	saksa	56,5	48,9	55,8	53,9	58,4
	20	aids	21,3	20,5	5,4	21,0	21,0
	24	japan	10,0	6,1	4,4	11,5	14,1
henk	1	summit	16,2	20,5	9,4	18,2	17,3
	12	bildt	42,1	38,1	34,1	42,1	74,3
	28	eyaho	11,8	8,7	22,7	22,6	18,2
	29	ydinv	64,9	38,6	46,6	67,1	93,2
org	5	varso	30,0	24,9	13,5	31,8	36,6
	8	opec	79,2	58,9	81,4	81,7	81,4
	11	eyval	20,3	20,2	18,6	20,2	24,9
	13	jugos	64,0	63,0	64,0	65,5	64,0
	15	valmet	48,6	50,2	48,6	50,4	48,6
	16	tampel	20,4	17,1	3,6	25,3	15,7
	17	matka	18,5	10,9	10,1	6,7	30,0
	18	neste	57,9	46,9	48,2	64,0	60,1
	30	vihr	32,1	16,0	33,1	17,7	37,3
paras			3	2	1	8	18

4.2 Relevantit

tyyppi	numero	aihe	perus	litt_fin	litt_tes	rak_fin	rak_tes
aihe	3	polku	19,2	45,8	33,2	47,4	50,0
	19	yjate	36,0	17,0	50,3	34,8	50,9
	21	elint	25,7	22,5	13,3	25,7	14,9
	22	asunt	5,0	4,9	5,0	5,1	5,0
	23	paast	42,6	30,1	32,6	45,2	43,2
	25	sellu	0,5	0,2	0,4	0,3	0,5
	26	aukio	12,3	9,1	23,6	10,5	31,0
	27	kierr	58,9	46,5	53,4	55,4	59,0
raj.aihe	2	velka	10,3	1,4	13,9	4,6	18,7
	4	jykul	78,8	68,4	45,4	0,3	72,8
	6	liett	67,3	67,3	31,5	67,3	69,7
	7	iraki	50,1	50,1	49,1	50,1	57,6
	9	bukar	48,8	46,9	49,0	48,1	59,9
	10	untag	68,1	68,0	25,8	68,8	74,2
	14	saksa	53,7	44,2	48,1	51,3	53,5
	20	aids	12,4	15,9	3,5	12,5	13,3
	24	japan	8,3	2,7	2,0	8,9	9,5
henk	1	summit	17,7	25,7	10,8	20,8	18,1
	12	bildt	52,1	48,3	32,4	52,1	80,6
	28	eyaho	11,0	11,8	17,9	17,8	17,1
	29	ydinv	16,3	14,2	17,0	23,3	42,8
org	5	varso	43,4	37,4	15,9	44,8	49,0
	8	opec	68,1	60,4	63,8	70,6	65,8
	11	eyval	12,0	12,0	12,9	12,0	14,8
	13	jugos	37,9	35,3	37,9	36,1	37,9
	15	valmet	40,5	42,5	40,5	44,5	40,5
	16	tampel	19,2	9,4	4,3	19,4	14,7
	17	matka	14,8	11,0	9,3	8,6	26,6
	18	neste	44,3	38,2	36,6	52,6	47,0
	30	vihr	34,9	13,8	24,8	22,1	28,4
paras			6	2	1	8	17

4.3 Erittäin relevantit

tyyppi	numero	aihe	perus	litt_fin	litt_tes	rak_fin	rak_tes
aihe	3	polku	20,7	54,9	43,8	54,6	60,6
	19	yjate	27,9	13,3	43,5	26,6	40,5
	21	elint	7,7	6,4	1,6	7,7	2,8
	22	asunt	3,5	3,5	3,5	3,5	3,5
	23	paast	21,8	13,6	20,1	22,0	22,3
	25	sellu	0,3	0,1	0,2	0,2	0,5
	26	aukio	5,9	5,5	11,7	5,1	14,7
	27	kierr	82,1	71,9	78,7	80,4	86,8
raj.aihe	2	velka	2,3	0,2	12,3	0,1	11,6
	4	jykul	77,1	68,8	33,0	0,3	66,5
	6	liett	17,3	17,3	7,1	17,3	16,5
	7	iraki	45,4	45,4	38,0	45,4	47,4
	9	bukar	32,2	30,7	33,7	32,2	50,1
	10	untag	29,6	29,8	9,4	30,1	31,7
	14	saksa	35,8	30,0	31,6	33,7	36,0
	20	aids	10,5	16,8	3,1	10,4	12,2
	24	japan	0,4	0,3	0,6	0,7	1,6
henk	1	summit	19,5	25,3	16,4	21,0	20,1
	12	bildt	3,9	2,9	3,1	3,9	4,6
	28	eyaho	7,1	2,5	7,3	7,5	13,1
	29	ydinv	3,1	2,3	4,6	7,1	14,0
org	5	varso	48,4	45,0	19,7	50,0	47,9
	8	opec	23,0	24,8	20,7	24,2	21,7
	11	eyval	6,0	6,0	13,3	6,0	9,4
	13	jugos	20,8	24,2	20,8	19,9	20,8
	15	valmet	4,8	4,5	4,8	4,9	4,8
	16	tampel	0,5	2,6	0,1	3,9	0,3
	17	matka	4,9	1,2	1,2	0,2	9,8
	18	neste	25,1	17,9	21,8	27,9	27,0
	30	vihr	12,5	8,6	6,9	50,3	9,6
paras			4	6	4	8	14