Perus- ja taivutusmuotohakemiston tuloksellisuus todennäköisyyksiin perustuvassa tiedonhakujärjestelmässä

Tuomas Kunttu

Informaatiotutkimuksen pro gradu -tutkielma
Marraskuu 2003
Informaatiotutkimuksen laitos
Tampereen yliopisto

Tampereen yliopisto
Informaatiotutkimuksen laitos
KUNTTU, TUOMAS: Perus- ja taivutusmuotohakemiston tuloksellisuus todennäköisyyksiin perustuvassa tiedonhakujärjestelmässä

Pro gradu -tutkielma, 80 s., 8 liites.
Informaatiotutkimus
Marraskuu 2003

TIIVISTELMÄ

Tutkimuksen tarkoituksena oli selvittää, miten tiedonhaun tuloksellisuus eroaa käytettäessä ositettua perusmuotohakemistoa, osittamatonta perusmuotohakemistoa ja taivutusmuotoista hakemistoa todennäköisyyksiin perustuvassa tiedonhakujärjestelmässä. Erityinen tutkimusongelma oli, kannattaako yhdyssanat osittaa hakemistossa. Tiedonhakujärjestelmänä oli Inquery-ohjelma ja aineistona informaatiotutkimuksen laitoksen suomenkielinen TUTK-tutkimustietokanta, jossa on noin 54 000 sanomalehtiartikkelia. Päätutkimusongelman selvittämiseksi 30 hakukysymyksestä muodostettiin perus- ja johdoskyselyt, jotka suoritettiin kaikkiin kolmeen hakemistoon. Tulokset laskettiin kolmella relevanssitasolla. Tiedonhaun tuloksellisuutta mitattiin laskemalla tarkkuuskeskiarvot vakioiduilla saantitasoilla sekä saanti ja tarkkuus dokumentin katkaisupisteittäin (DCV).

Tulosten mukaan ositettu perusmuotoistettu hakemisto on tuloksellisin hakemisto. Hieman huonommin menestyivät osittamaton perusmuotohakemisto ja taivutusmuotohakemisto. Näiden kahden välillä ei juuri ollut eroja. Tulokset olivat samansuuntaiset kaikilla kolmella tutkitulla relevanssitasoilla.

Kokotekstitietokantojen, joiden sisällön kielenä on morfologialtaan rikas ja runsaasti yhdyssanoja sisältävä kieli, hakemisto kannattaa perusmuotoistaa ja yhdyssanat osittaa, koska 1) hakujen tuloksellisuus on parempi 2) tiedonhakijan on helpompi hakea sanojen perusmuodoilla.

SISÄLLYS

1 Johdanto

Tiedonhakuun liittyvä kielitieteellinen tutkimus on ollut suurelta osin englannin kieltä koskevaa. Näiden tutkimuksien tuloksia ei kuitenkaan usein voi verrata suomenkielisiin tiedonhakujärjestelmiin, koska suomen kieli poikkeaa etenkin morfologialtaan suuresti englannista. Tässä tutkimuksessa halutaan selvittää, kannattaako suomenkielisessä tietokannassa käyttää perus- vai taivutusmuotoista hakemistoa, kun tiedonhakujärjestelmän täsmäytys perustuu todennäköisyyksien laskemiseen. Erityisesti kiinnostuksen kohteena on yhdyssanojen osituksen hyödyllisyys. Alkula (2000) on tutkinut hakemistojen eroja täydellisesti täsmäyttävässä Boolen järjestelmässä, mutta osittaistäsmäyttävissä järjestelmissä ei asiaa ei ole tutkittu. Tämä tutkimus pyrkii täyttämään tältä osin aukon tiedonhakututkimuksen kentässä. Koska tietokannan erilaisia hakemistoja verrataan, on suomen kielen morfologiaan kiinnitettävä erityistä huomiota.

Tutkimuksen tarkoituksena on vertailla kahden hakutavan tuloksellisuutta perusmuotohakemiston, jossa yhdyssanat on ositettu, toisen perusmuotohakemiston, jossa yhdyssanoja ei ole ositettu sekä taivutusmuotohakemiston välillä todennäköisyyksiin perustuvassa Inquery-tiedonhakujärjestelmässä. Vertailu suoritetaan tekemällä kaikkiin hakemistoihin peruskyselyt ja johdoksilla laajennetut johdoskyselyt. Tutkimusympäristönä on informaatiotutkimuksen laitoksen TUTK-tutkimustietokanta, jossa on noin 54 000 sanomalehtiartikkelia kokotekstinä. Perusmuotoiset hakemistot ovat entuudestaan olemassa ja tutkimusta varten on luotu taivutusmuotoinen hakemisto.

Osaongelmana on, tuottavatko peruskyselyt vai loogisesti muodostetut johdoskyselyt parempia tuloksia hakemistojen sisällä. Toisena osaongelmana tutkitaan, voidaanko probabilistisessa järjestelmässä taivutusmuotoisessa hakemistossa katkaisua simuloida seulontamenetelmällä eli suorittamalla haku kaikilla hakemistossa esiintyvillä taivutusmuodoilla. Lisäksi tutkitaan, mitkä ovat eri hakemistojen heikkoudet ja vahvuudet.

Tutkimustietokannalle on olemassa 35 hakuaihetta, joille relevanssiarviot on tehty. Tässä tutkimuksessa kyselyt suoritetaan 30 hakuaiheesta perus- ja taivutusmuotohakemistoon. Luonteeltaan tutkimus on laboratorioympäristössä toteutettava evaluointitutkimus.

Luvut kahdesta neljään ovat teoriaosuutta. Toisessa luvussa käydään läpi tiedonhaun käsitteet ja seuraavassa luvussa kielitieteen käsitteet. Neljännessä luvussa katsotaan, mitä ongelmia luonnollinen kieli aiheuttaa tiedonhaussa ja miten ongelmia on ratkaistu. Tutkimuksen kulku selvitetään viidennessä luvussa. Ensin kerrotaan tutkimusympäristöstä, sitten hakuaiheista, kyselyiden muodostamisesta ja tilastollisesta testauksesta. Tulokset käsitellään ensin koko kyselyjoukon osalta ja sitten hakuaiheiden käsitetyypeittäin luvussa kuusi. Tulosluvun lopussa etsitään syitä hakemistojen erilaisiin tuloksiin sekä lasketaan kyselyiden sanamääriä ja yhdyssanojen osuuksia. Luvussa seitsemän on keskusteluosuus, jossa tuloksia pohditaan ja suhteutetaan aiempaan tutkimukseen. Johtopäätökset ovat viimeisessä kahdeksannessa luvussa. Liitteessä yksi on lueteltu hakukysymykset ja liitteessä kaksi ovat hakulauseet.

2 Tiedonhaku

2.1 Tiedonhaun käsitteet

Tiedonhaku kohdistuu **tietokantaan**, joka voidaan määritellä "järjestetyksi yhtenäiseksi tietorakenteeksi tiedon hakua, selaamista, käsittelyä (ylläpitoa) varten" (Tietotekniikan sanasto 1990, 488). Fidelin (1987, 5) mukaan tietokanta on kokoelma tietoa, joka on valikoitu todellisesta maailmasta ja jota käytetään määriteltyihin tarkoituksiin. Tietokanta muodostuu **tietueista**, jotka kuvaavat yhtä rajattua yksikköä (henkilöä, kirjaa, uutista). Rakenteiset tietueet puolestaan muodostuvat **kentistä**, joissa kussakin kuvataan yleensä yhtä tietueen yksikön ominaisuutta (henkilön tai kirjan nimeä, uutistekstiä tms.).

Nykyään tekstitietokantojen tiedostorakenne perustuu useimmiten **käänteisrakenteeseen** (inverted file structure), koska sieltä hakeminen on huomattavasti nopeampaa kuin silloin, kun käytetään tietokantaa ilman käänteisrakennetta. Tällöin hakua suoritettaessa kone joutuisi käymään tietokannan jokaisen tietueen läpi yksi kerrallaan. Käänteisrakenteen nopeus korostuu varsinkin silloin, kun yhdistetään monta hakuavainta samaan kyselyyn. Käänteisrakenteinen tietokanta muodostuu tyypillisesti vähintään kolmesta tiedostosta: varsinaisesta tietokannasta, käänteistiedostosta ja käänteistiedoston hakemistosta. **Käänteistiedostoon** (inverted file) poimitaan kaikki tietueissa esiintyneet hakuavaimet, eli haettavissa kentissä esiintyvät sanat. Hakuavaimen yhteyteen tulee tieto, missä tietueissa ja kentissä kyseinen sana sijaitsee. Tarkempikin tieto hakuavaimen sijainnista on mahdollinen. Mikäli käänteistiedostossa esitetään tieto monennessako lauseessa ja monentenako sanana hakuavain on, voidaan haussa käyttää läheisyysoperaattoreita. **Sanakirjatiedosto** (dictionary file) on käänteistiedoston hakemisto. Sinne tulee vain tieto siitä, monessako tietueessa on kutakin hakuavainta käytetty. Se voi sisältää tiedon hakuavaimen sijainnista käänteistiedostossa. Suhteellisen pienen kokonsa vuoksi sanakirjatiedostoa on nopea käyttää. (Järvelin 1995, 94–98.) Käänteistiedostosta on käytetty tässä tutkimuksessa myös sanaa **hakemisto**.

Tässä tutkimuksessa käytetään sanaa **perusmuotohakemisto** tietokannan käänteistiedostosta, jonka hakuavaimet on perusmuotoistettu Twol-ohjelmalla. **Taivutusmuoto-**

hakemisto puolestaan on käänteistiedosto, jossa hakuavaimet esiintyvät taivutusmuotoisina eli siinä muodossa kuin ne varsinaisessa tietokannassa ovat.

Järvelin (1995, 68–73 ja 176–178) on esittänyt tiedonhaun tasoperiaatteen. Tiedonhaun hakutehtävä ja dokumentit voidaan esittää kolmella tasolla: käsite-, ilmaisu- ja esiintymätasolla. **Käsitetasolla** hakutehtävän tai dokumentin aihetta analysoidaan miettien, mitä käsitteitä se sisältää. Tulokseksi saadaan kuvaus dokumentin käsitteellisestä sisällöstä ja voidaan laatia käsitteellinen hakusuunnitelma. **Ilmaisutasolla** näille löydetyille käsitteille haetaan vastineet joko luonnollisesta kielestä tai jostain dokumentaatiokielestä. **Hakuavain** on yleisnimitys hakutermeille, hakusanoille sekä muille koodeille ja lyhenteille. Hakuavain vastaa ilmaisutasolla käsitetason hakukäsitteitä. **Hakutermit** ovat dokumentaatiokielen tai jonkin muun erityiskielen termejä. **Hakusanat** ovat luonnollisen kielen sanoja tai ilmauksia. Tiedonhaku tapahtuu varsinaisesti aina **esiintymätasolla**, koska tietokoneet käsittelevät dataa tällä tasolla. Edellisen tason hakuavaimet ovat tällä tasolla **merkkijonoja**. Nämä jakautuvat merkkijonokaavioihin ja merkkijonovakioihin. **Merkkijonovakiot** ovat kokonaisia merkkijonoja, jotka vastaavat ilmaisutason hakuavaimia. **Merkkijonokaavio** on osa merkkijonovakiota. Se vastaa siis ilmaisutason katkaistua hakuavainta. Kuten Järvelin (1995, 187) muistuttaa, esiintymätason merkkijonovakiot ovat hyvin läheisessä suhteessa ilmaisutason hakuavaimiin. Niinpä tässä tutkimuksessa puhutaan aina hakuavaimista, riippumatta kumman tason ilmiöistä on kyse.

Kun hakuavaimia yhdistetään erilaisilla operaattoreilla yhteen, saadaan **hakulausekkeita**. Näitä hakulausekkeita voidaan kutsua myös **kyselyiksi**. **Hakukysymyksiksi** kutsutaan tässä tutkimuksessa tiedontarvitsijan muotoilemia tiedontarpeita, jotka hän esittää tiedonhakijalle. Hakukysymys voi siis esimerkiksi olla:

OPEC:n öljyn hintaa ja tuotantomääriä koskevat päätökset.

Jonka perusteella tehty kysely perusmuotohakemistoon voisi olla:

#sum(opec öljy hinta tuotantomäärä koskea päätös)

2.2 Tiedonhaun lähestymistavat

Tiedonhakua voidaan lähestyä eri suunnista. Järvelin (1995, 30–34) esittelee neljä merkittävää lähestymistapaa.

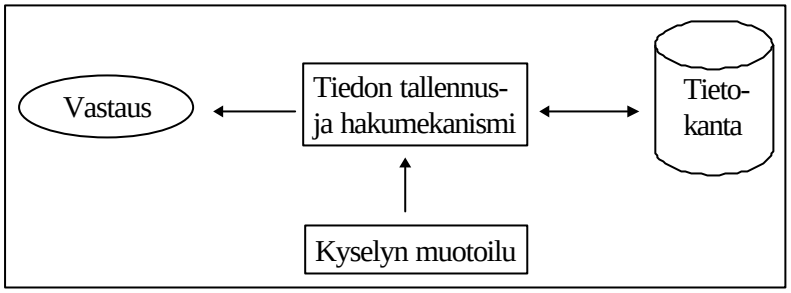
1. Tarkasteltaessa **tiedonhakua täsmäyttämisenä** voidaan kaikissa hakujärjestelmissä nähdä kolme osaa: kyselyt, dokumenttijoukon ja näiden välissä täsmäytysmekanismin, joka määrää mitkä dokumentit täsmäävät kyselyyn (Salton & McGill 1984, 10–11).

2. **Tiedonhaku teknisenä prosessina** erittelee käytännössä ne prosessit, joita tiedonhakuun liittyy. Siinä esitetään dokumenttien tuotanto, hankinta ja tallennus sekä toisaalta tiedontarpeen muotoutuminen ja sen muotoilu kyselyksi. Nämä kaksi puolta yhdistyvät tiedon hakumekanismissa.

3. **Kognitiivinen näkökulma** esittelee tiedonhakua tiedon tuottajan, välittäjän ja käyttäjän tiedollisten prosessien kautta. Se selittää, miten tietämys siirtyy tiedon tuottajalta välittäjän kautta käyttäjälle ja miten nämä tietämysrakenteet vaikuttavat toisiinsa. (Ingwersen 1992, 135.)

4. Tiedonhaun tuloksellisuus ja kustannukset ovat keskiössä **evaluoivassa näkökulmassa**. Tarkastelu voi tapahtua makro- tai mikrotasolla. Makroevaluoinnissa huomio kiinnittyy tiedonhaun kokonaisprosessin tuloksiin, kuten saadun tiedon laatuun tai määrään. Mikrotasolla tarkastellaan hakuprosessin eri tekijöitä ja vaiheita.

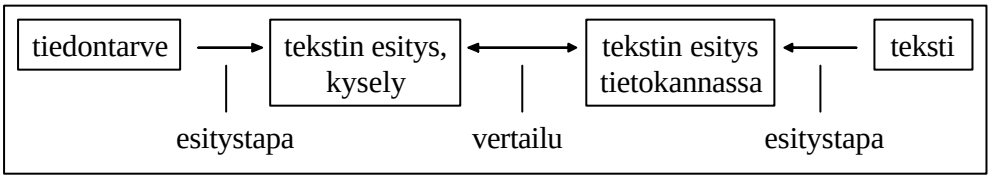
Tämä tutkimus on tiedonhaun evaluointitutkimus, joten tässä kiinnostavin näistä on evaluoiva näkökulma (ks. kuvio 1). Tiedonhaun prosessia ajatellen evaluointi tässä tutkimuksessa tapahtuu kapealla alueella. Koska kyse on erilaisten hakujen tuloksellisuuden vertailusta erilaisissa käänteistiedostoissa, toimitaan siis tiedonhakumekanismin, tietokannan ja kyselyn muotoilun tasolla. Tutkimuksen tarkastelu tapahtuu siis mikrotasolla.



KUVIO 1. Tiedonhakuprosessin evaluointikehys (muokattu Järvelin 1995, 33).

2.3 Tiedonhakumenetelmät

Tiedonhaun asetelma voidaan esittää kuvion 2 kuvaamalla tavalla. Tiedontarve muotoillaan kyselyksi eli kysely on tiedontarpeen esitys. Toisaalla itse tekstiä edustaa tietokannassa sen esitys. Kyselyä ja tekstin esitystä verrataan toisiinsa ja mitä enemmän ne toisiaan muistuttavat, sitä paremmin ne täsmäävät. (Belkin & Croft 1987, 110.)

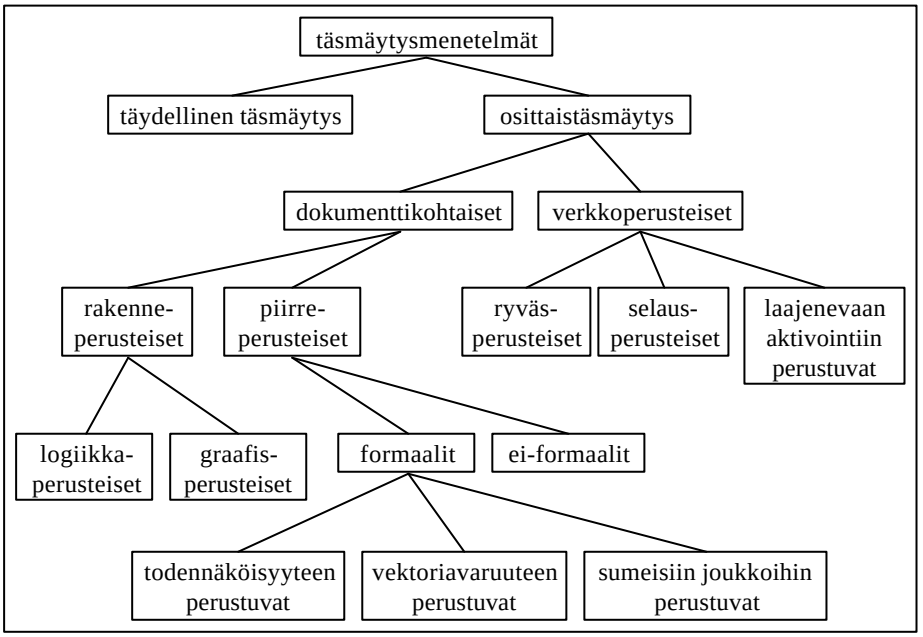


KUVIO 2. Tiedonhaun asetelma (Belkin & Croft 1987, 110).

Erilaisia tiedonhakumenetelmiä voidaan erotella täsmäytyksen perusteella. Belkin ja Croft (1987, 112) jaottelevat tiedonhakutekniikat kuvion 3 mukaisesti. Pääjako on täys- ja osittaistäsmäytykseen. Täystäsmäytyksessä, joka perustuu Boolen logiikkaan, kyselyssä esiintyvien hakuavainten on esiinnyttävä dokumenteissa paitsi samanmuotoisina myös juuri annettuna loogisena kombinaationa. Vaikka kaupallisissa järjestelmissä täystäsmäytys on ollut laajassa käytössä, siinä on kuitenkin monia ongelmia. Belkinin ja Croftin (1987, 113) mukaan täystäsmäytyksen menetelmille on ominaista, että

1. ne eivät löydä relevantteja dokumentteja, joihin kysely täsmää vain osittain
2. ne eivät lajittele tuloksia relevanssijärjestykseen
3. avaimia ei voi painottaa kyselyissä eikä dokumenteissa
4. hakulausekkeen looginen muotoilu on mutkikasta.

Osittaistäsmäytykseen perustuvat menetelmät ovat pyrkineet ratkaisemaan täystäsmäytykseen liittyneitä ongelmia ja ovatkin ratkaisseet kaikki yllä mainitut täystäsmäytykseen liittyvät ongelmat. Osittaistäsmäytys on hajautunut useisiin menetelmiin. Ne voidaan jakaa ensinnäkin dokumenttikohtaisiin menetelmiin, jotka nimensä mukaisesti vertaavat kyselyä yksittäisen dokumentin korvikkeeseen ja verkkoperusteisiin menetelmiin, jotka painottavat yksittäisen dokumentin sijaan dokumenttien verkkoa, yhteyksiä dokumenteista toisiin dokumentteihin. Dokumenttikohtaiset menetelmät jaetaan edelleen rakenne- ja piirreperusteisiin menetelmiin. Rakenneperusteisissa menetelmissä täsmäytys perustuu johonkin formaaliin logiikkaan tai dokumenteista ja kyselyistä tehtyjen kuvioiden vertailuun. Piirreperusteisissa menetelmissä kyselyt ja dokumentit esitetään piirrejoukkoina, esimerkiksi indeksointiavaimina. Formaaleihin menetelmiin kuuluvat eniten tutkitut osittaistäsmäyttävät menetelmät: todennäköisyyteen, vektoriavaruuteen ja sumeisiin joukkoihin perustuvat menetelmät. (Emt., 112–113.)



KUVIO 3. Täsmäytysmenetelmät Belkinin ja Croftin (1987, 112) mukaan.

Näistä formaaleista menetelmistä etenkin vektorimalli ja todennäköisyysmalli ovat olleet ahkeran tutkimuksen kohteina (Kekäläinen 1999, 21). Vektorimallissa dokumentit ja kysely esitetään moniulotteisen avaruuden vektoreina. Kukin dokumenttivektori sisältää kaikki tietokannan sisältämät sanat, mutta painotettuina. Jos sana kuvaa dokumenttia hyvin, se saa arvon 1 ja jos ei lainkaan niin arvo on 0. Myös arvot ykkösen ja nollan välistä ovat käytössä. Kyselyvektori muodostetaan samalla periaatteella. Tärkeimmille

käsitteille voidaan antaa suurempia painoja kuin muille käsitteille. Kyselyvektorissa on yhtä monta komponenttia kuin dokumenttivektorissa. Näiden vektorien samankaltaisuutta voidaan tutkia erilaisin matemaattisin menetelmin, yleisimmin käytetään kosinifunktiota. Yleisesti voidaan sanoa, että mitä pienempi on vektorien välinen kulma, sitä paremmin kyseessä oleva dokumentti täsmää kyselyyn. (Järvelin 1995, 121–124.)

2.3.1 Todennäköisyysmalli

Todennäköisyysmallin kehittely on alkanut 1960-luvulla, mutta 1970-loppupuolelta lähtien on julkaisujen määrä aiheesta kasvanut. Yksi usein viitatuista on todennäköisyyslajitteluperiaatteen kehittäjä S.E. Robertson (1977). Tämän jälkeen todennäköisyysmallia ovat käsitelleet mm. van Rijsbergen (1979), Bookstein (1984) ja Cooper (1994)

Tiedonhaun todennäköisyysmallin lähtökohta on todennäköisyyslajitteluperiaate, joka Cooperin (1994, 242) mukaan määritellään seuraavasti:

> "Systeemin yleinen tehokkuus käyttäjälle on käytettävissä oleviin tietoihin nähden paras, kun hakusysteemi tuottaa vastauksena jokaiseen kyselyyn tietokannan kaikki dokumentit järjestettynä käyttäjän kannalta alenevan käyttökelpoisuuden todennäköisyyden mukaan. Nämä todennäköisyydet on laskettu mahdollisimman tarkasti kaiken saatavilla olevan tiedon avulla."[1]

Toisin kuin täystäsmäytyksessä dokumentin relevanssi ei ole dikotominen, vaan pikemminkin suhteellinen. Todennäköisyyslajitteluperiaatteen (probabilibty ranking principle) mukaan dokumenttien relevanssille lasketaan todennäköisyyksiä ja tulosjoukko järjestetään näiden mukaan.

[1] "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data is available for this purpose, then the overall effectiveness of the system to its users will be the best obtainable of the basis of that data."

Todennäköisyyteen perustuvan tiedonhaun perusmalli on binaarinen riippumattomuusmalli. Se olettaa, että dokumenteissa on toisistaan riippumattomat merkkijonot ja niillä paino 1, jos merkkijono esiintyy tai 0, jos merkkijono ei esiinny.

Jokaista dokumenttia kuvaa binaarinen vektori $x = (x_1, x_2, \ldots x_n)$, jossa $x_i = 0$ tai 1 riippuen siitä, onko dokumentissa kyseinen hakuavain. Todennäköisyyslajitteluperiaatteen optimaalinen lajittelufunktio laskee dokumentin relevanssin seuraavasti:

$$\frac{P(R \mid D)}{P(NR \mid D)} \qquad (1)$$

Tässä P(R | D) tarkoittaa ehdollista todennäköisyyttä, että havaittu dokumentti D on relevantti ja P(NR | D) tarkoittaa todennäköisyyttä, että havaittu dokumentti D ei ole relevantti. Käyttämällä Bayesin muunnossääntöä saadaan lajittelufunktio sopivampaan muotoon

$$\frac{P(D \mid R) * P(R)}{P(D \mid NR) * P(NR)} \qquad (2)$$

Lajittelufunktion kaavassa P(D | R) voidaan estimoida avainten esiintymisen tai puuttumisen perusteella seuraavasti:

$$P(D \mid R) = \prod_{i=1}^{t} (p_i)^{x_i} (1 - p_i)^{1-x_i} \qquad (3)$$

jossa

P(D | R) tarkoittaa todennäköisyyttä, että havaitaan D, kun havaitaan relevantti dokumentti

p_i on todennäköisyys, että hakuavain i:llä on paino 1 relevanttien dokumenttien joukossa eli

$p_i = P(x_i = 1 \mid rel)$

$(1 - p_i) = P(x_i = 0 \mid rel)$

$x_i = 1$, jos hakuavain i esiintyy dokumenteissa, muuten $x_i = 0$

i = hakuavain, jotka käydään läpi ensimmäisestä viimeiseen eli 1…t

Kaavassa lasketaan siis termien $(p_i)^{x_i}$ ja $(1-p_i)^{1-x_i}$ tulo, kun i käy 1:stä t:hen.

Vastaavasti voidaan estimoida P(D | NR) seuraavasti:

$$P(D \mid NR) = \prod_{i=1}^{t} (q_i)^{x_i} (1 - q_i)^{1-x_i} \qquad (4)$$

jossa

P(D | NR) tarkoittaa todennäköisyyttä, että havaitaan D, kun havaitaan ei-relevantti dokumentti

q_i on todennäköisyys, että hakuavain i:llä on paino 1 ei-relevanttien dokumenttien joukossa eli

$q_i = P(x_i = 1 \mid \text{non-rel})$

$(1 - q_i) = P(x_i = 0 \mid \text{non-rel})$

Kun P(D | R) ja P(D | NR) sijoitetaan kaavaan ja niistä otetaan logaritmit, saadaan lineaarinen erottelufunktio

$$g(D) = \sum_{i=1}^{t} x_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_{i=1}^{t} \log \frac{1-p_i}{1-q_i} + \log \frac{P(R)}{R(NR)} \qquad (5)$$

Summan termeistä vain ensimmäinen vaikuttaa dokumenttien relevanssijärjestykseen. Niinpä lauseke voidaan kirjoittaa muotoon

$$g(D) = \sum_{i=1}^{t} x_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad (6)$$

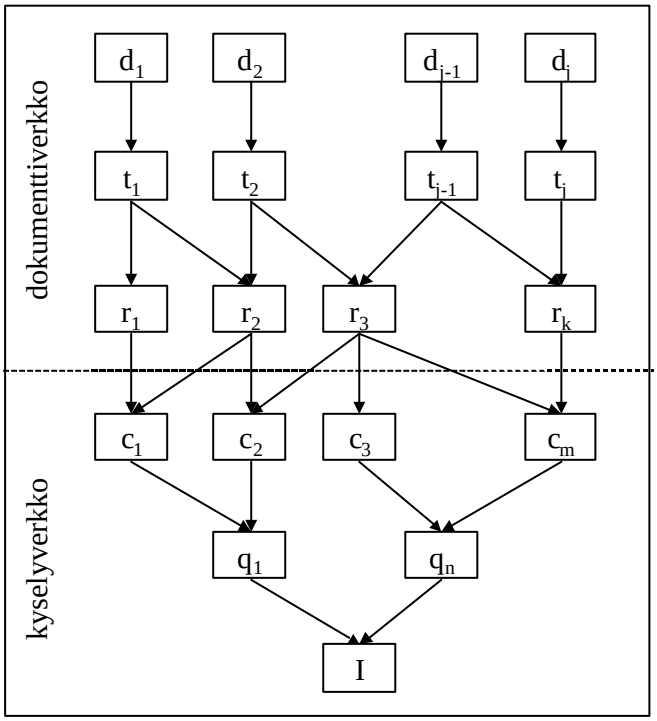
Tällä kaavalla saadaan siis relevanssilajitteluarvo dokumentille D kyselyn q suhteen. (Kekäläinen 1999, 22–24; Croft & Harper 1979, 286–287.)

Muutamat tutkijat ovat kritisoineet tiedonhaun todennäköisyysmallia. Yhteenvedon kritiikistä on esittänyt mm. Cooper (1994, 243). Todennäköisyyslajitteluperiaate olettaa, että dokumentin hyödyllisyys käyttäjälle on binaarinen. Mutta todellisessa tilanteessa se on tietysti asteittainen täysin hyödyttömästä erittäin hyödylliseen. Myös todennäköisyyslajitteluperiaatteen yleistä pätevyyttä on kritisoitu. Hyödyllisyyden todennäköisyyden mukaan tulosjoukossa järjestetyt dokumentit eivät aina ole parhaassa mahdolli-

sessa järjestyksessä, vaikka todennäköisyydet olisikin laskettu mahdollisimman tarkasti. Joissakin tilanteissa dokumentit voitaisiin järjestää paremminkin. Kowalskin (1997, 99) mukaan ongelmat johtuvat siitä, että matemaattista mallia sovellettaessa puuttuu täsmällistä dataa ja lisäksi yksinkertaistetaan oletuksia. Hän huomauttaa kuitenkin, että nämä todennäköisyysmallin heikkoudet voidaan kääntää eduiksi, koska ne tunnetaan hyvin ja niiden korjaamisen eteen voidaan tehdä töitä.

2.4 Inquery-hakujärjestelmä

Inquery-hakujärjestelmä on kehitetty Massachusettsin yliopistossa. Se perustuu dokumenttihaun päättelyverkkoon (document retrieval inference network), joka on eräs Bayesin päättelyverkon sovellus. Nämä päättelyverkot perustuvat tiedonhaun todennäköisyysmalliin. Bayesin päättelyverkko on suunnattu, syklitön verkko, jossa solmukohdat edustavat propositionaalisia muuttujia ja niitä yhdistävät viivat riippuvuussuhteita. Solmukohdat muodostavat puumaisen mallin, ja kunkin solmukohdan arvo on joko tosi (true) tai epätosi (false) riippuen edeltävistä solmukohdista. Päättelyverkko muodostuu kahdesta osasta: dokumentti- ja kyselyverkosta. (Callan, Croft & Harding 1992.)



KUVIO 4. Päättelyverkko (Turtle 1991, 39).

Turtlen (1991, 38–46) esittelemän päättelyverkon (kuvio 4) perusmallin solmut ovat viidellä tasolla, minkä lisäksi kyselyverkossa I tarkoittaa tiedontarvetta. Dokumenttiverkossa jokainen dokumenttisolmu (d) edustaa todellista dokumenttia tietokannassa. Dokumenttien tekstejä edustavat tekstisolmut (t). Tekstisolmun riippuvuus dokumenttisolmusta esitetään päättelyverkossa näitä yhdistävällä viivalla. Sisällön esityssolmut (r, content representation node) edustavat tekstejä. Dokumenttiverkko rakennetaan vain kerran kullekin tietokannalle. Kyselyverkon puolella kyselysolmut (q) edustavat tiedontarvetta (I). Kyselyn käsitesolmut (c, query concept node) edustavat kyselyissä olevia käsitteitä. Kyselyverkko muodostetaan jokaiselle kyselylle uudestaan, ja se muuttuu kun kyselyä uudelleenmuotoillaan. Verkot kohtaavat, kun r- ja c-solmujen vastaavuudet lasketaan. Turtlen mukaan usein perusmallista käytetään yksinkertaistettua mallia, jossa oletetaan, että dokumentti- ja tekstisolmut ja toisaalta kyselysolmut ja kyselyn käsitesolmut vastaavat aina toisiaan. Näin yksinkertaistetussa mallissa teksti- ja kyselyn käsitesolmut jätetään pois. (Turtle 1991, 38–46.) Inqueryssa käytetään juuri tällaista yksinkertaistettua mallia (Callan ym. 1992).

Inqueryn tärkeimmät tehtävät ovat dokumentti- ja kyselyverkon luominen ja tämän jälkeen päättelyverkoston käyttäminen tiedonhakuun. Dokumenttiverkkoa luotaessa jokainen dokumentti muunnetaan joukoksi sisällön esityssolmuja. Tätä kutsutaan dokumenttien jäsentämiseksi (parsing). Jäsentäminen koostuu viidestä vaiheesta: leksikaalinen analyysi, syntaktinen analyysi, käsitetunnistus, käsitteiden tallennus ja toimintojen kirjaaminen (transaction generation). 1) Leksikaalisessa analyysissä dokumenteista tunnistetaan sanat, numerot ja kenttätunnisteet, muutetaan teksti pienaakkosiksi sekä poistetaan mahdolliset sulkusanat. Myös pääteaineksen karsinta (stemmaus) voidaan suorittaa tässä vaiheessa. 2) Syntaktisessa analyysissä varmistetaan, että dokumentti on oikeassa muodossa ja sitä voidaan korjata, jos näin ei ole. 3) Käsitetunnistuksessa voidaan mm. tunnistaa kirjaimin kirjoitettuja numeroita ja muuttaa ne numeroiksi. Myös yritysten ja henkilöiden nimien tunnistaminen on mahdollista. 4) Kun dokumentin sanat indeksoidaan sanakirjatiedostoon, korvataan merkkijono numeerisella tunnisteella, koska näin säästetään tallennustilaa. Myöhemmin samaan sanaan viitataan aina samalla numerolla. Tämä vaihe on käsitteiden tallennus. 5) Joka kerta, kun termi tunnistetaan, sen sijainti dokumentissa kirjautuu muistiin. Kun dokumentin kaikki termit saadaan käsiteltyä, tiedostossa ovat indeksointitapahtumat, jotka sisältävät kunkin termin frekvenssit ja sijainnit kyseessä olevassa dokumentissa. Jokainen toiminto (transaction) edustaa linkkiä

dokumentin solmun ja sisällön esityssolmun välillä. Koko dokumenttiverkkoa edustaa joukko toimintotiedostoja (transactions files), jotka syntyivät jäsennyksessä. Dokumentin jäsennyksen jälkeen toimintojen kirjauksessa tallennetut indeksointitiedot järjestetään ja niistä muodostetaan käänteistiedosto. Se sisältää tiedot termien esiintymismääristä tietokannassa ja dokumenttien määrän, joissa termi esiintyy sekä tiedot toiminnoista, joissa termi esiintyi. (Callan ym. 1992.) Informaatiotutkimuksen laitoksella käytössä oleva Inquery-ohjelmassa ei ole käytössä käsitetunnistusta.

Inqueryssa haun alajärjestelmä (retrieval subsystem) muuntaa kyselyn kyselyverkoksi ja vertaa sitä jo olemassa olevaan dokumenttiverkkoon. Kyselyä käsiteltäessä se muutetaan pienaakkosiksi, siitä voidaan poistaa sulkusanat ja jäljelle jääneistä sanoista voidaan poistaa pääteaines. Ohjelman sanakirjaa käytetään apuna käsitteiden tunnistamisessa. Sitten Inqueryn hakukone vertaa kyselyverkkoa ja dokumenttiverkkoa toisiinsa ja palauttaa kyselyn mukaan määritellyt todennäköisyydet, joiden mukaan tulosjoukko järjestetään. Todennäköisyydet ovat arvio tiedontarpeen täyttämisestä. (Emt.)

2.4.1 Inqueryn käyttämä painotusfunktio

Monien muiden tiedonhakujärjestelmien tapaan Inquery käyttää tf*idf-kaavaan arvioitaessa dokumentin relevanssin todennäköisyyttä. Tf tarkoittaa avainten frekvenssiä dokumentissa ja idf käänteistä dokumenttifrekvenssiä, joka kertoo monessako dokumentissa hakuavain esiintyy. Dokumenttifrekvenssin käänteisyys saa aikaan sen, että mitä harvemmissa dokumenteissa termi esiintyy mahdollisimman monta kertaa, sitä korkeampi on dokumentin relevanssin todennäköisyys. Toisin kuin monet muut järjestelmät Inquery aloittaa oletustodennäköisyydestä ja korjaa arviota evidenssin perusteella (Broglio, Callan, Croft & Nachbar 1995). Inqueryn versiossa 3.1 käytetään seuraavaa kaavaa laskettaessa painotusfunktio sanalle t:

$$0.4+0.6*\left(\frac{tf_{td}}{tf_{td}+0.5+1.5*\frac{length(d)}{avglen}}\right)*\left(\frac{\log\frac{N+0.5}{n_t}}{\log N+1}\right) \qquad (7)$$

jossa

tf_{td} = avaimen t esiintymisfrekvenssi dokumentissa d

length(d) = dokumentin d pituus

avg len = tietokannan dokumenttien keskimääräinen pituus

N = dokumenttien lukumäärä tietokannassa

n_t = dokumenttien määrä, jotka sisältävät avaimen t (Allan ym. 1998).

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

Tällä kaavalla saadut painot ovat välillä [0, 1], joten ne voidaan tulkita avainten esiintymisen todennäköisyyksiksi dokumentissa.

2.4.2 Inqueryn operaattorit

Inqueryn operaattorit alkavat aina #-merkillä. Operaattorien jälkeen tulevat suluissa hakuavaimet tai kyselyn alilausekkeet, joita kyselyssä halutaan käyttää. Tämän työn kannalta tärkeimmät operaattorit ovat #sum, #syn sekä #odN . Inqueryn ohjeiden (1996) mukaan summaoperaattori #sum painottaa kaikkia hakuavaimia yhtä paljon. Se laskee avainten painoarvoista keskiarvon ja näin saadaan #sum-solmun painoarvo. Operaattorin #sum tuottama painoarvo saadaan kaavasta

$$\frac{(p_1 + p_2 + ... + p_n)}{n} \qquad (8)$$

jossa

p = #sum-operaattorin sisältämä hakuavain tai -lauseke

n = hakuavainten tai -lausekkeiden määrä (Callan ym. 1992).

Synonyymioperaattorin #syn yhdistämiä hakuavaimia kohdellaan kuten saman avaimen esiintymiä. Synonyymioperaattori muuttaa Inqueryn painotusfunktion seuraavaksi (Kekäläinen 1999, 25).

$$0.4 + 0.6 * \left(\frac{\sum_{t \in S} tf_{td}}{\sum_{t \in S} tf_{td} + 0.5 + 1.5 * \frac{length(d)}{avglen}} \right) * \left(\frac{\log \frac{N + 0.5}{n_t}}{\log N + 1} \right) \qquad (9)$$

jossa
S = syn-operaattorin hakuavainten joukko
Muut symbolit on selitetty kaavassa 7.

Operaattori #odN (ordered distance N) tai vain #N on läheisyysoperaattori, jossa N määrittää, kuinka kaukana sanat saavat olla toisistaan. Operaattori edellyttää, että kaikki sen sisältämät hakuavaimet esiintyvät tietueessa annetussa järjestyksessä. (Inquery document retrieval system 1996.) Esimerkkinä olkoon lause:

	Tämä	*lause*	*on*	*pieni*	*esimerkki.*
sanan sijaluku:	1	2	3	4	5

Esimerkissä sanojen *lause* ja *esimerkki* välinen etäisyys on 5–2=3, joten hakulause *#3(lause esimerkki)* täsmäisi, mutta *#2(lause esimerkki)* ei.

2.5 Tiedonhaun evaluointi

Tiedonhaun tutkimuksen yksi keskeisiä tutkimuskohteita on tiedonhaun evaluointi. Uuden hakualgoritmin suorituskykyä täytyy tietysti aina verrata vanhoihin tiedonhaun menetelmiin. Millä perusteilla sitten määritellään tiedonhakujärjestelmien paremmuus? Evaluointikriteerit voidaan jakaa kahtia kustannus- ja laatukriteereihin. Kustannuskriteerejä ovat mm. käytön hinta, käytön helppous, vastausaika ja tulosten käytettävyys. Laatukriteereitä ovat aineiston kattavuus, saanti ja tarkkuus, uutuus ja virheettömyys. Nämä voidaan myös yhdistää laskemalla kustannusten ja laadun suhde. Kustannuskriteerejä etenkin kustannus-hyöty-ajatteluna voidaan kutsua myös **tehokkuudeksi** eli järjestelmän kyvyksi suoriutua tehtävästään mahdollisimman edullisella kustannus-hyöty-suhteella. Laatukriteereitä voi kutsua **tuloksellisuudeksi** eli järjestelmän kyvyksi tehdä sitä, mihin se on tarkoitettu. Tässä tutkimuksessa tiedonhakua evaluoidaan tuloksellisuuden kannalta, erityisesti keskittyen saantiin ja tarkkuuteen. Kuitenkin ennen kuin niitä pystytään mittaamaan, täytyy selvittää relevanssin käsite.

2.5.1 Relevanssi

Relevanssi on informaatiotutkimuksen keskeisimpiä termejä. Onhan tiedonhaun tarkoituksena juuri löytää relevantteja dokumentteja tiedontarvitsijalle. Siitä huolimatta, tai ehkä juuri sen vuoksi, relevanssin määrittely on tuottanut ongelmia.

Relevanssi on perinteisesti jaettu aihe- ja käyttäjärelevanssiin. Aiherelevanssi tarkoittaa yksinkertaisimmillaan kyselyn ja dokumentin täsmäyttämistä. Mitattavuutensa vuoksi sitä on pidetty helpoimpana relevanssin määritelmänä. Käyttäjärelevanssissa tiedontarvitsija määrittää relevanssin asteen sen perusteella, miten käyttökelpoiseksi hän dokumentin kokee. Käyttäjärelevanssin määrittelyssä on tuotu esiin sellaisia käsitteitä kuin käyttökelpoisuus, hyödyllisyys, tilannekohtainen relevanssi ja tyytyväisyys. (Järvelin 1995, 43–44.)

Relevanssin määrittelemiseksi on kehitetty erilaisia malleja. Saracevic (1975) esittelee laajasti erilaisia relevanssin määritelmiä. Myöhemmin relevanssia ovat käsitelleet mm. Saracevic (1996) sekä Cosijn ja Ingwersen (2000). Katsauksen relevanssin historiaan on viimeksi esittänyt Mizzaro (1997). Saracevic (1996, 214) on esittänyt relevanssille viisi ilmenemismuotoa:

- **Systeemi- tai algoritmirelevanssi** on kyselyn ja informaatio-objektin välinen suhde.
- **Aiherelevanssi** on kyselyssä ilmenevän aiheen ja informaatio-objektin sisältämän aiheen välinen suhde.
- **Kognitiivinen relevanssi** tai **asiaankuuluvuus** on tiedonkäyttäjän tietämyksen tilan tai kognitiivisen tilan ja haettujen informaatio-objektien välinen suhde.
- **Tilannerelevanssi** tai **hyödyllisyys** kuvaa tilanteen, tehtävän tai ongelman ja haetun informaatio-objektin välistä suhdetta.
- **Motivaatio-** tai **affektiivinen relevanssi** on käyttäjän tavoitteiden, aikomusten tai motivaation ja haetun informaatio-objektin välinen suhde.

Cosijn ja Ingwersen (2000) tarkastelivat Saracevicin esittämiä relevanssin ominaisuuksia. He myös poistivat ilmenemismuodoista motivaatiorelevanssin, koska heidän mukaan motivaatiorelevanssi ja affektiivinen relevanssi ovat kaksi erillistä relevanssityyp-

piä, joista motivaatiorelevanssi sisältyy aikomus-ominaisuuteen ja affektiivinen relevanssi liittyy kaikkiin subjektiivisiin relevanssityyppeihin. Cosijn & Ingwersen lisäsivät relevanssin ilmenemismuotoihin hyvin kontekstiriippuvaisen sosio-kognitiivisen relevanssin, joka liittyy mm. tiedeyhteisön vuorovaikutukseen. (Emt., 549.)

Tässä tutkimuksessa käytetty relevanssin määrittely on aiherelevanssia. Se on relevanssin muodoista ainoa, joka sopii tiedonhaun laboratoriotutkimuksiin (Kekäläinen 1999, 94).

2.5.2 Saanti ja tarkkuus

Tiedonhaun tuloksellisuuden mittaamiseen on kehitetty useita mittareita, mutta saanti ja tarkkuus ovat kaikkein useimmin käytetyt. **Saanti** (recall) on haussa saatujen relevanttien dokumenttien osuus kaikista relevanteista dokumenteista. Se osoittaa siis systeemin kykyä esittää kaikki relevantit dokumentit. **Tarkkuus** (precision) puolestaan on relevanttien dokumenttien osuus kaikista löydetyistä dokumenteista. Tarkkuus siis osoittaa systeemin kykyä esittää vain relevantit dokumentit. Saanti ja tarkkuus ilmaistaan tavallisesti desimaaliluvuilla [0, 1] tai prosentteina 0 %–100 %.

Näiden mittareiden taustalla on ajatus, että keskimääräinen käyttäjä haluaa saada suuren määrän relevantteja dokumentteja (saanti) ja samalla mahdollisimman vähän epäolennaisia dokumentteja (tarkkuus). Näinhän ei aina ole. Varsinkin saantia tiedonhaun tuloksellisuuden mittarina on kritisoitu, koska se ilmeisen huonosti sopii yhteen tiedonhaun hyötyteoreettisen lähestymistapaan, joka on perusta monille tiedonhaun teorioille. (Salton 1992, 442.) Käyttäjälle saattaa riittää tiedontarpeen täyttämiseen yksi ainoa relevantti dokumentti ja se voi olla tulosjoukon kärjessä, vaikka saanti muuten olisi huono. Absoluuttisen saannin laskeminen edellyttää, että tietokannan kaikki dokumentit käydään läpi, jotta tiedetään, mitkä dokumentit ovat relevantteja kuhunkin kyselyyn.

Vuorovaikutteisen tiedonhaun käyttäjätutkimuksessa (Su 1992, 512) havaittiin, että haun onnistumisella (success) ja tarkkuudella ei ollut merkitsevää yhteyttä käyttäjien kannalta. Yhteyttä ei myöskään ollut tarkkuuden ja käyttäjän tyytyväisyyden välillä. Tutkimuksen mukaan tarkkuus ei ollutkaan hyvä mittari tiedonhaun tuloksellisuutta mittaaman, saantia arvostettiin enemmän (emt., 514).

Joka tapauksessa saanti ja tarkkuus ovat ylivoimaisesti eniten käytettyjä, helposti tulkittavia ja tarkkuuden osalta myös helposti laskettavia tiedonhaun tuloksellisuuden mittareita.

On selvää, että saanti ja tarkkuus ovat toisistaan riippuvaisia. On havaittu, että saannin noustessa tarkkuus laskee ja päinvastoin. Cleverdon (1972, 199) muotoili asian näin: "Yleiseksi säännöksi jää, että riittävästi toistettuna, saannin paraneminen voidaan saavuttaa vain tarkkuuden heikkenemisen kustannuksella." Hän myös muistuttaa, että kyseessä ei kuitenkaan ole informaatiotutkimuksen peruslaki. Saannin ja tarkkuuden käänteistä suhdetta on selitetty mm. indeksoinnin tyhjentävyydellä ja spesifisyydellä ja toisaalta indeksoinnin johdonmukaisuuteen ja indeksointitermien epätäsmällisyyteen liittyvillä seikoilla (Järvelin 1995, 57).

Yleinen tapa havainnollistaa hakujen tuloksellisuutta on laatia saanti- ja tarkkuusarvojen perusteella saanti-tarkkuus-käyrä. Siitä nähdään, miten tarkkuus muuttuu saannin kasvaessa. Osittaistäsmäyttävissä järjestelmissä tarkkuus voidaan laskea vakioiduille saantitasoille. Tällöin evaluointi perustuu yhtäläiseen suoritustasoon. Tämä kuvaa tiedonhaun tuloksellisuutta järjestelmän näkökulmasta. Tulosten tulkinnan ongelmana on kuitenkin, että saantikantojen koko – eli kunkin kyselyn relevanttien dokumenttien määrä – vaihtelee paljonkin hakupyynnöittäin. Saantikanta saattaa olla vain muutaman dokumentin suuruinen, toisinaan siihen voi kuulua satoja dokumentteja. (Hull 1993.)

Käyttäjät ovat usein kiinnostuneita lähinnä hakutuloksen kärjestä, niistähän tulosten selaaminen aloitetaan, ja kovin montaa dokumenttia käyttäjät eivät välttämättä jaksa käydä läpi. Tätä voidaan tutkia laskemalla saanti ja tarkkuus tietyn suuruisille tulosjoukoille eli tietyissä katkaisukohdissa. Näistä tehdyt kuviot ovat DCV (document cut-off value) -käyriä. Katkaisupisteet voivat olla esimerkiksi 1, 10, 20, 30, tällöin siis lasketaan, mikä on saanti ja tarkkuus, kun yksi dokumentti on löydetty, kun 10 dokumenttia on löydetty jne. Nämä käyrät kuvaavat hyvin sitä, mitä käyttäjät saavat hakutuloksen kärkeen. Tämä tapa kuvaa siis tiedonhakijan näkemää vaivaa. DCV-käyrien ongelma on saantikantojen koon vaihtelu. Tarkkuus hakutuloksessa tuloksen koolla 30 ei voi olla korkea, jos saantikannan koko on viisi dokumenttia. (Hull 1993.)

3 Luonnollinen kieli

3.1 Kielen osajärjestelmät

Luonnolliset kielet ovat monitasoisia, monimutkaisia ja joustavia järjestelmiä. Monitasoisia, koska niissä voidaan erottaa useita osajärjestelmiä. Monimutkaisia, koska kielen hallitsemiseksi tarvitaan paljon sääntöjä, joita niistä tehdyt kieliopit yrittävät kuvata. Kuitenkin kielet ovat joustavia järjestelmiä, koska niiden avulla pystytään ilmaisemaan lähes kaikkea, mitä ihminen ajattelee ja tuntee.

Kielen perustavat osajärjestelmät Karlssonin (1998, 15–16) mukaan ovat: semantiikka, syntaksi, leksikko, morfologia ja fonologia. **Semantiikka** eli merkitysoppi tutkii kielessä (sanoissa, lauseissa ja kieliopillisissa kategorioissa) olevia merkityksiä. **Syntaksin** eli lauseopin tutkimuskohde on lauseiden rakenne. Syntaksi tutkii, mistä lauseet rakentuvat, mitä tehtäviä sen eri osasilla on ja miten näitä osasia voi yhdistää. **Leksikko** eli sanasto sisältää kielen vakiintuneet sanat. **Morfologia** eli muoto-oppi tutkii sanojen sisäistä rakennetta ja esimerkiksi sanojen taivutusta ja johtamista. **Fonologian** tutkimuskohteena ovat kielten äännejärjestelmät. Morfologia on tämän tutkimuksen kannalta tärkein kielen osajärjestelmä. Myös semanttisia ilmiöitä liittyy tähän tutkimukseen. Näitä kielen osajärjestelmiä käsitellään seuraavissa kappaleissa enemmän.

3.2 Morfologia

Sana on yksi kielen peruskäsitteistä. Puheessa ja kirjoituksessa esiintyvät sanat ovat **sananmuotoja** eli saneita. Suomen kielessä sanat taipuvat. Sanoilla on **perusmuoto** ja **taivutusmuotoja**. Yhdessä nämä muodostavat sanan **taivutusparadigman**. Perusmuodon ja taivutusmuotojen yhteinen osa on sanan **vartalo**, johon taivutusmuodoissa lisätään päätteitä. Vartalo voi olla selvästi näkyvissä, kuten *talo – talo+i+ssa*, tai siinä voi tapahtua erilaisia morfofonologisia vaihteluita, kuten *punainen – punaise+ssa* tai *yö – ö+i+nä*. Taivutuksen perusteella sanat voidaan jakaa nomineihin, jotka taipuvat sijoissa, verbeihin, jotka taipuvat persoonissa, tempuksissa ja rajoitetusti sijoissa sekä partikkeleihin, jotka eivät yleensä taivu lainkaan. Nominien perusmuotona pidetään yksikön nominatiivimuotoa ja verbeillä 1. infinitiivin lyhyempää muotoa.

Morfeemi on kielessä pienin merkitystä kantava yksikkö. Morfologisessa analyysissä sanoista eli lekseemeistä erotellaan morfeemit segmentoimalla. Morfeemit voivat olla vapaita tai sidonnaisia. **Vapaat morfeemit** voivat esiintyä yksinään, ilman että siihen liittyy muita morfeemeja. **Sidonnaiset morfeemit** eivät voi esiintyä ilman toista morfeemia. Näitä ovat esimerkiksi **affiksit** eli liitteet, joka on yleisnimitys mm. **prefikseille** eli etuliitteille ja **suffikseille** eli loppuliitteille.

3.2.1 Morfeemien jaottelu

Morfien eristäminen onnistuu, kun morfin rajat ovat rakenteellisesti ja semanttisesti läpinäkyviä. Nämä ovat **agglutinatiivisia** sanarakenteita, joissa päätteitä laitetaan peräkkäin (*talo+ssa+ni+kin*). Hankalampaa on, kun kaksi merkitysyksikköä on sulautunut toisiinsa. Sitä kutsutaan **fuusioksi**. Tästä esimerkkinä ovat germaanisten kielten vahvat verbit (ruotsissa *skina – sken, dra – drog*). Näitä imperfektimuotoja pidetäänkin yhtenä morfina ja niitä kutsutaan **salkkumorfeiksi**, koska ne pitävät yhtä aikaa sisällään sekä sanan merkityksen että menneen ajan. (Karlsson 1998, 92.)

Morfianalyysi voi paljastaa aineksia, joilla ei ole selvää merkitystä tai funktiota ja eivät siksi ole tyypillisiä morfeja. Nämä ovat **jäännösmorfeja**. Suomessa näitä ovat esimerkiksi *van+hurskas, vento+vieras, puti+puhdas, ruti+köyhä*. Jäännösmorfit ovat usein kielen historiallisen kehityksen myötä hämärtyneitä jäänteitä. Morfien erikoistyyppejä ovat myös **nollamorfit**, jotka eivät toteudu foneemeina tai grafeemeina. Ne eivät siis tule esille kirjoituksessa eikä äännettäessä. Suomessa tällaisia ovat substantiivien yksikön nominatiivit ja verbien preesensit (*auto, sano+n*). (Emt., 94.)

Allomorfit ovat morfeemien variantteja, muunnoksia. Saman morfeemin allomorfit ovat rakenteeltaan osittain samanlaisia mutta merkitykseltään identtisiä. Suomessa on kolmenlaisia allomorfeja. Näistä kaksi ensimmäistä liittyy morfofonologiseen vaihteluun. Vokaalisoinnun aiheuttama vaihtelu suomen päätteissä tuottaa allomorfeja. Esimerkiksi morfeemi *–ssA*, joka ilmaisee inessiiviä, toteutuu allomorfeina *–ssa* ja *–ssä*. Toiseksi sanan vartalo voi vaihdella. Esimerkiksi morfeemilla *kalA*, on kaksi allomorfia: *kala-* ja *kalo-*. Kolmanneksi kaksi morfia voi olla saman morfeemin allomorfeja, jos ne ovat muodoltaan erilaisia, mutta merkitykseltään samanlaisia. Tästä esimerkkinä on 3. persoonan omistusliitteet *auto+ssa+an* ja *auto+ssa+nsa*. (Emt., 94–95.)

3.2.2 Kielten morfologinen typologia

Jokaisella kielellä on omat sääntönsä siitä, miten morfeemit liittyvät toisiinsa. Karlsson (1998, 116–118) esittelee neljä kielten morfologista päätyyppiä. **Isoloivissa kielissä** sanat eivät taivu. Lauseiden sanojen väliset suhteet ilmaistaan sanajärjestyksellä, partikkeleilla tai prosodisilla keinoilla eli äänteiden kestolla, painotuksella tai intonaatiolla. Isoloivia kieliä ovat tyypillisesti vietnam ja kiina. **Agglutinoivissa kielissä** sanoihin kiinni liitetään morfeemeja ilmaisemaan sanojen taipumista. Morfofonologista vaihtelua ei ole. Turkki sekä monet uralilaiset kielet ovat agglutinoivia kieliä. Suomikin on lähinnä agglutinoiva kieli. Esimerkiksi sana auto taipuu agglutinoivasti: *auto – auto+n – auto+i+ssa*. Tosin suomessa on paljon morfofonologista vaihtelua: *käsi – käde+n – kät+tä.* **Fuusioivissa kielissä** on vartalon sisäistä äännevaihtelua, kuten englannin laulaa-verbin taivutus: *sing – sang – sung*. Sanat eivät taivu affikseilla, kuten agglutinoivissa kielissä, vaan erillisillä partikkeleilla. Fuusioivia kieliä ovat esimerkiksi germaaniset kielet. Neljäntenä tyyppinä ovat **polysynteettiset kielet**. Näissä kielissä on sidonnaisia morfeemeja hyvin runsaasti ja niiden merkitys on voimakkaampi kuin tavanomaisten päätteiden. Niinpä muissa ryhmissä lausetta vastaava rakennelma on polysynteettisissä kielissä sanan näköinen. Tällaisia kieliä ovat mm. eskimo- sekä intiaanikielet.

3.2.3 Suomen kielen morfologiaa

Suomen kieli on siis morfologisessa typologiassa lähinnä agglutinoiva kieli, jossa on kuitenkin myös fuusiokielten ominaisuuksia, koska sanavartalossa tapahtuu paljon morfofonologisia muutoksia. Toisaalta suomi on synteettinen kieli, mikä tarkoittaa, että suomen kielessä on paljon morfeemeja sanojen määrää kohti. Suomelle on ominaista, että vartalomorfeemin perään voi liittyä paljon sidonnaisia morfeemeja. Näitä suffiksimorfeemeja on neljänlaisia.

1. **Johdin** on suffiksimorfeemi, jolla johdetaan olemassa olevista sanoista uusia sanoja, esimerkiksi *kahvi+**la***.
2. **Tunnus** on myös suffiksimorfeemi. Tunnuksia ovat nomineilla monikon tunnus, adjektiiveilla komparatiivin ja superlatiivin tunnus sekä verbeillä mm.

aikamuodon ja tapaluokan tunnukset, *esimerkiksi talo+**i**+ssa, iso+**mpi**, istu+**i**, tul+**isi**+n*.

3. **Päätteellä** ilmaistaan sanan suhdetta muuhun lauseeseen. Päätteitä suomessa ovat vain nomineihin liittyvät sijapäätteet sekä verbien persoonapäätteet, esimerkiksi *talo+**ssa,** tule+**n***.

4. **Liitteet** ovat omistusliitteitä eli possessiivisuffikseja tai liitepartikkeleita, esimerkiksi *talo+**ni**, hän+**kin***. (Leino 1991, 41–42.)

Näiden suffiksimorfeemien järjestys suomessa käy ilmi esimerkiksi sanasta *lato+mo+i+ssa+mme+kin*. Sanavartalon jälkeen ensin tulee johdin, sitten tunnus, pääte, omistusliite ja viimeiseksi liitepartikkeli. Lähimpänä vartaloa on siis suffiksi, joka eniten muuttaa sanavartalon merkitystä.

3.2.4 Uusien sanojen muodostaminen

Sanojen johtaminen on suomessa tärkein tapa muodostaa kieleen uusia sanoja. Myös tässä tutkimuksessa sanojen johtaminen on keskeisessä asemassa, koska peruskyselyistä muodostettiin johdoskyselyt.

Johtaminen tapahtuu tyypillisesti siten, että kielessä jo olemassa olevaan sanaan, **kantasanaan** lisätään **johdin**, jolloin saadaan **johdos** eli uusi sana. Johtosuhde osoittaa kantasanan ja johdoksen välisen suhteen. Sitä on tapana merkitä kulmamerkin avulla. Siis esimerkiksi *ilo > iloinen*. Johdoksia on suomen kielen sanoista noin 44 % (Lepäsmaa, Lieko & Silfverberg 1996, 12). Erilaisia johtimia on suomen kielessä parisataa. Eniten on nomineihin lisättäviä **denominaaleja** johtimia. Melko paljon on myös verbikantaisia **deverbaaleja** johtimia. Myös jostain partikkeleista voidaan muodostaa johdoksia. Korostettakoon tässä, että niin denominaali-, deverbaali- kuin partikkelikantaisten johtimien tuottama uusi johdos voi olla niin nomini, verbi kuin partikkelikin.

Aina ei johtaminen kuitenkaan ole yksinkertainen ja selkeä kantasana + johdin = johdos -prosessi. Joskus johtaminen voi olla **takaperoista** eli kantasanalta näyttävä sana onkin johdos ja päinvastoin. Tällainen on esimerkiksi *rieha < riehaantua*, jossa ensin on ollut olemassa sana *riehaantua*, josta on johdettu sana *rieha*. Kielessä on paljon sanoja, joille ei voida osoittaa selvää kantasanaa. Tällaisissa tapauksissa kantasana on saattanut hävitä

aikojen kuluessa tai kyseessä voi olla **korrelaatiojohtaminen**. Siinä sanoja johdetaan käyttämällä mallina jonkin toisen sanan johtosuhdetta. Esimerkiksi johtosuhteesta *sairas* > *sairastaa* > *sairastua* on otettu mallia johdettaessa *viisas*-sanaa: *viisas* > *Ø* > *viisastua*. Näin syntyy aukollisia johdosketjuja. (Lepäsmaa ym. 1996, 15–18.)

Johtosuhteiden selvittäminen ei siis ole mikään mekaaninen prosessi, vaan vaatii myös sanojen morfologian tuntemisen lisäksi etymologista tietoa. Esimerkiksi tutkimusaineistossa olleen *maa*-sanan johdokseksi voisi kuvitella *maata*-verbin. Mutta sanojen etymologiat kertovat, että *maa* on ikivanha omaperäinen sana, kun taas *maata* on ilmeisesti germaanista perua oleva lainasana eikä siten voi olla *maa*-sanan johdos (Suomen sanojen alkuperä 1995, 133–136).

Johdokset voivat olla läpinäkyviä eli **transparentteja** tai läpinäkymättömiä eli **opaakkeja**. Transparentit johdokset voidaan helposti hahmottaa johdoksiksi, sillä niissä on näkyvissä kantasana ja johdin, kuten sanoista *viro* > *virolainen*. Opaakit johdokset eivät helposti hahmotu johdoksiksi. Tällaiset johdokset ovat **leksikaalistuneet** eli johdokset ovat siirtyneet leksikkoon ja näin muuttuneet perussanoiksi. (Kieli ja sen kieliopit 1994, 219.) Näistä esimerkkinä sanat *käsi* > *käsittää*.

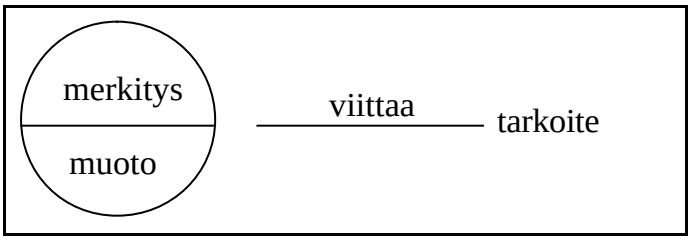
Johtimien **produktiivisuus**, eli se pystytäänkö niillä edelleenkin muodostamaan uusia sanoja, vaihtelee. Hyvin produktiivisia ovat mm. tekijäjohdin *-jA (tekijä, kalastaja*) ja verbistä substantiivin tekevä -minen (*tekeminen, kalastaminen*). Sen sijaan esimerkiksi aikaa ilmaisevalla *-oin*-johtimella ei enää uusia sanoja muodosteta *(piakkoin, hiljakkoin, jolloin*). Kirjallisuudesta löytyy ainakin kaksi produktiivisten johtimien luetteloa. Vesikansa (1977) esittelee kirjassaan noin 140 johdinta, joista mainitsee erittäin produktiiveiksi 12 ja produktiiveiksi niin ikään 12. Lepäsmaa ym. (1996) esittelee 65 johdinta, joista nimeää 13 erittäin produktiiveiksi ja 25 produktiiveiksi.

Sanojen yhdistäminen on toinen tärkeä tapa luoda uusia sanoja kieleen. Yhdyssanojen alkuosa on **määriteosa** ja jälkiosa on **perusosaa**. Yhdyssana voi olla **rinnasteinen**, jolloin sanan osat ovat samanarvoisia, kuten *suomalais-ruotsalainen* tai nykyään tuskin edes yhdyssanaksi hahmottumaton *maailma*. **Alisteisessa** yhdyssanassa määriteosa kuvaa tarkemmin perusosaa, kuten *luonnonsuojelu*. Yhdyssanoja on kielessä runsaasti. Lepäsmaan ym. (1996, 12) mukaan niitä on suomessa noin 44 % sanoista. Iisan, Oitti-

sen ja Piehlin (1999, 117) mukaan Nykysuomen sanakirjassa niitä on 65 % sanoista. Johtamisen ja yhdistämisen lisäksi sanoja tulee kieleen lainaamalla ja lyhentämällä.

3.3 Semantiikka

Semantiikka jaetaan sana- ja lausesemantiikkaan. Tässä yhteydessä on tarpeen keskittyä sanasemantiikkaan. Jokaisella sanalla eli lekseemillä on jokin merkitys, käsite. Toisaalta kaikilla sanoilla on myös muoto, joka on primaaristi ääntä eli sana puhuttuna. Mutta monilla sanoilla on muotona myös kirjoitusasu. Merkityksen ja muodon suhde on konventionaalinen. Tämä tarkoittaa sitä, että suhde perustuu sosiaaliseen sopimukseen. On siis sovittu, että tietyt äänteet peräkkäin tarkoittavat tiettyä käsittä. Yhdessä sanan merkitys ja muoto viittaavat johonkin reaalimaailman tarkoitteeseen. (Ks. kuvio 5.) Sanan merkitys sisältää ne ominaisuudet, jotka reaalimaailman tarkoitteen tulee täyttää, jotta viittaaminen juuri siihen tarkoitteeseen sanalla olisi mahdollista. (Karlson 1998, 12–13.)



KUVIO 5. Muodon ja merkityksen suhde tarkoitteeseen (muokattu Karlsson 1998, 12).

Sanoilla on erilaisia merkityksiä. Ne voivat vaihdella käyttäjän ja tilanteen mukaan. Sanan merkitys voidaankin jakaa denotatiiviseen ja konnotatiiviseen merkitykseen. **Denotatiivinen merkitys** on sanan päämerkitys, kielenpuhujille yhteinen merkitys. Päämerkitys rajaa sanan mahdollisen tarkoitteen. **Konnotatiivinen merkitys** on sivumerkitys. Siihen sisältyy käyttäjän kokemuksiin perustuvia merkityksiä. Konnotatiivinen merkitys ei voi enää muuttaa tarkoitetta, vaan tarkentaa sitä. (Karlson 1998, 235.)

Synonymia on tilanne, jossa eri sanoilla on sama merkitys, mutta eri muoto. Synonyymeiksi kelpaavat vain itsenäiset sanat, joiden muoto on riittävän erilainen. Tiukimmillaan synonyymien pitäisi olla vaihdettavissa kaikissa konteksteissa ilman, että merkityksessä tapahtuu mitään muutosta. Täydellistä synonymiaa esiintyy kielissä harvoin; usein on kyse lähisynonymiasta. Näitä ovat mm. ruveta ~ ryhtyä ~ alkaa tai koira ~ rak-

ki. (Karlson 1998, 219–220.) Monimerkityksisillä sanoilla on enemmän kuin yksi denotaatio. **Homonymia ja polysemia** ovat ilmiöitä, joissa eri sanoilla on sama muoto, mutta eri merkitys. Muotojen samanlaisuus voi olla täydellistä, siis kaikissa taivutusmuodoissa esiintyvää, tai riittää myös, että vain perusmuodot ovat samanlaisia. Sen sijaan agglutinoivissa kielissä, kuten suomessa, usein esiintyvä taivutusmuotojen samanlaisuus ei ole varsinaista homonymiaa. Esimerkkinä taivutusmuotohomonymiasta on sanan-muoto *hauista*, joka palautuu perusmuotoihin *haku, hauki* ja *hauis*. Homonymian ja polysemian ero on sanojen alkuperissä. Jos sanoilla on eri alkuperä, eli ne ovat kaksi eri sanaa, jotka vain sattuvat olemaan samanmuotoisia, kyseessä on homonymia. Tästä esimerkkinä sana *kuusi*, joka voi tarkoittaa havupuuta tai numeroa 6. Polysemiassa sanoilla on sama alkuperä, josta – usein metaforan kautta – ovat merkitykset eriytyneet. Esimerkiksi sanat *laskea (mäkeä, lukuja)* ja *selkä (ihmisen, järven, kirjan)* ovat polyseemisia. (Karlson 1998, 213–214.)

4 Luonnollinen kieli tiedonhaussa

Tiedonhaussa ollaan aina tekemisissä sekä luonnollisen että keinotekoisen kielen kanssa. Keinotekoinen kieli liittyy hakulauseiden muotoilussa operaattorien käyttöön. Luonnollista kieltä käytetään hakuavaimissa sekä tietueissa, joihin hakuavaimet täsmäytyvät. Luonnollisen kielen moninaisuudesta ja joustavuudesta johtuen törmätään tiedonhaussa usein kielestä johtuviin ongelmiin.

Tämän luvun ensimmäisessä alaluvussa pureudutaan luonnollisen kielen aiheuttamiin ongelmiin tiedonhaussa. Luvussa 4.2 tarkastellaan Alkulan (2000) tutkimusta hakujen tuloksellisuudesta perus- ja taivutusmuotohakemistossa. Kolmannessa alaluvussa esitellään tämän tutkimuksen kannalta keskeiset luonnollisen kielen analysointiohjelmat.

4.1 Luonnollisen kielen aiheuttamia ongelmia ja niiden ratkaisuja tiedonhaussa

Pirkola (1999, 42–43) on esittänyt viisi kieleen liittyvää ongelmaa tiedonhaussa.

1. **Vaihtoehtoisten käsitteiden ja hakuavainten valinta**. Samaa aihetta käsiteltäessä siitä voidaan käyttää useita erilaisia sanoja. Tiedonhakujärjestelmän käyttäjät eivät pysty tarpeeksi tyhjentävästi esittämään kaikkia käsitettä kuvaavia sanoja.
2. **Hakuavainten morfologinen vaihtelu**. Sanojen taipuminen, johtaminen ja yhdisteleminen vaikeuttavat täsmäytymistä.
3. **Hakuavainten viittaussuhteet ja poisjättämiset** liittyvät anaforiin ja elliptisiin ilmauksiin.
4. **Hakuavainten monitulkintaisuus** liittyy sanojen homonymiaan ja polysemiaan.
5. **Monikielisyys** aiheuttaa ongelmia. Haku suomenkielisillä sanoilla englanninkielisestä tietokannasta ei tietenkään onnistu ilman hyvin kehittynyttä tiedonhakujärjestelmää.

Luonnollisen kielen käsittelyä eli NLP:tä (natural language processing) on käytetty ratkaisemaan kielen tiedonhakuun aiheuttamia ongelmia. NLP:n tarkoituksena on saada tietokoneet käsittelemään luonnollista kieltä. Tekstit analysoidaan automaattisesti, jotta

tiedonhaku, automaattinen kääntäminen tai muu tekstin käsittely onnistuisi (Haas 1996, 83).

Tämän tutkimuksen kannalta kiinnostavin ongelma liittyy hakuavainten morfologiseen vaihteluun. Näitä ongelmia on pyritty ratkaisemaan mm. hakuavainten katkaisulla (truncation), perusmuotoistamisella (normalization), pääteaineksen karsinnalla (stemming) sekä tuottamalla taivutusvartalot tai kaikki taivutusmuodot (Pirkola 1999, 47–48; Alkula 2000, 84–89).

Merkkijonokatkaisussa hakuavain katkaistaan manuaalisesti sopivasta kohtaa katkaisumerkillä (esimerkiksi * tai ?). Tämä katkaistu hakuavain on merkkijonokaavio. Tällöin haku täsmäytyy kaikkiin kyseisellä merkkijonolla alkaviin sanoihin eli merkkijonovakioihin. Esimerkiksi hakuavain *luon** täsmäytyy sanoihin *luonto, luonnossa, luontoarvo, luonne, luonteenpiirr*e jne. Jos katkaistu hakuavain on kovin lyhyt, haku täsmäytyy hyvin moniin sanoihin ja hakutulos voi olla suuri. Kaikilla sanoilla ei yhteistä vartaloa ole ollenkaan, kuten *yö – öiden*. Merkkijonokatkaisu edellyttää tiedonhakijalta kielen morfologian tuntemusta, jotta hän osaa katkaista sanat sopivista kohdista.

Sanojen **perusmuotoistaminen** tapahtuu kielen analysointiohjelman avulla. Ohjelma muuttaa kaikki taivutusmuotoiset sanat perusmuodoiksi. Tietokannan käänteistiedosto voidaan perusmuotoistaa, jolloin saadaan perusmuotohakemisto. Tällöin hakuavaimet voidaan syöttää kyselyyn niin ikään perusmuodossa. Perusmuotoistamisen ongelmat liittyvät suomen kielen taivutusmuotohomonymiaan. Kun taivutusmuoto *hauista* muutetaan kaikkiin kolmeen perusmuotoon (haku, hauki, hauis), voidaan hakuavaimella *hauki* saada tiedonhaun kirja. Toinen perusmuotoistamisen ongelma liittyy perusmuotoistamisohjelmien "liian huolelliseen" toimintaan (Alkula 2000, 107). Ne saattavat tuottaa täysin teoreettisia perusmuotoja, kuten sananmuodon *kokkolasta* perusmuodot *kokkola* ja *kokkolapsi* tai sananmuodon *kuin* perusmuodot *kuin* ja *kuu*. Näistä *kokkolapsi* ja *kuu* perusmuototulkinnat ovat kyllä morfologisesti oikein, mutta semanttisesti ne eivät ole mielekkäitä. Kolmas ongelma liittyy perusmuotoistamisohjelman sanakirjaan. Se ei ole koskaan täydellinen, ja sieltä puuttuvia sanoja ohjelma ei tietenkään pysty perusmuotoistamaan. Kun tämän puuttuvan sanan perusmuodolla sitten suoritetaan haku, täsmäytystä ei tapahdu. Perusmuotoistamisesta aiheutuvia monitulkintaisuusongelmia voidaan ratkaista mm. sanaluokkadisambiguoinnilla, jossa analyysiohjelma tunnistaa sanojen

sanaluokat automaattisesti tekstiyhteyden perusteella. Tämä auttaa tietenkin vain niissä tapauksissa, jossa sanojen sanaluokka on eri, kuten *kuusi* (numero tai puu) tai *voin* (perusmuoto *voi* tai *voida*).

Pääteaineksen karsinnassa eli stemmauksessa sanoista poistetaan ohjelmallisesti liitteet, päätteet, tunnukset ja joskus johtimetkin, jolloin jäljelle jää sanan vartalo. Karsinta voidaan suorittaa mahdollisimman täydellisesti, mutta tällöin vartalo voi jäädä hyvin lyhyeksi ja ongelmat ovat samoja kuin katkaisussa. Tunnetuin stemmeri on Martin Porterin (1980) esittelemä Porter stemmer.

Morfologisen tulkintaohjelman avulla voidaan tuottaa sanan kaikki **taivutusvartalot**. Sana syötetään ohjelmaan perusmuodossa ja ohjelma tuottaa taivutusvartalot. Esimerkiksi *kauppa*-sanasta taivutusvartalot ovat *kauppa, kaupa, kauppoi, kauppoj* ja *kaupoi*. Sitten haku voidaan suorittaa taivutusvartaloilla katkaisun avulla. Näin hakutulokseen tulevat varmasti kaikki kauppa-sanan taivutusmuodot, mutta vältetään pitkälti tavallisen katkaisun ongelmat, koska tavallisella katkaisulla hakuavain olisi *kaup** ja hakutulokseen tulisivat mm. kaikki *kaupunki*-alkuiset sanat.

Ohjelmallisesti voidaan tuottaa myös halutun sanan kaikki **taivutusmuodot**. Sitten kaikki taivutusmuodot voidaan lisätä hakulauseeseen. Tämä tapa sopii kielille, joissa on yksinkertainen morfologia eli sanat taipuvat vain muutamissa sijoissa. Esimerkiksi ohjelmaan syötettävä englannin sana *mouse* tuottaa taivutusmuodot *mouse, mouse's, mice* ja *mice's*. Voimakkaasti taipuvien kielten tiedonhakuun, kuten suomeen, kaikki taivutusmuodot tuottava ohjelma ei sovellu. Voihan suomen substantiivilla olla jopa 2000 taivutusmuotoa.

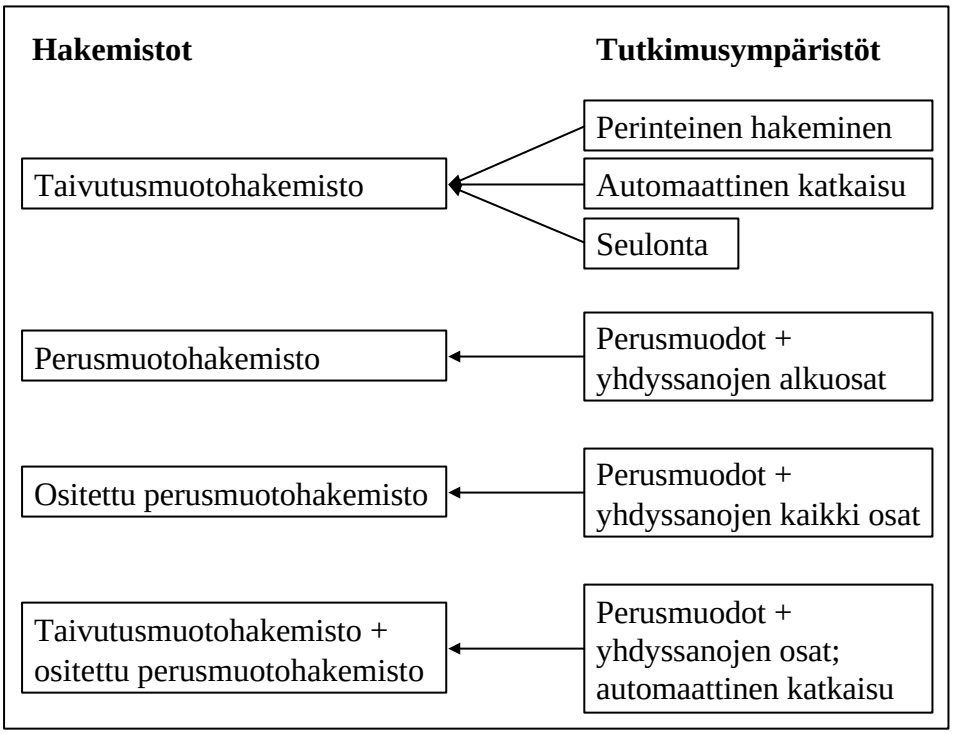
4.2 Hakemistojen tuloksellisuuden vertailu täystäsmäyttävässä järjestelmässä

Hakujen tuloksellisuutta erimuotoisista hakemistoista on tutkittu aiemmin vain kerran Alkulan (2000) tutkimuksessa. Hän on selvittänyt, miten suomen kielen morfologisten tulkintaohjelmien avulla voidaan ratkaista tiedonhakuun liittyviä ongelmia. Tiedonhakujärjestelmänä käytettiin täystäsmäytykseen perustuvaa Basis-järjestelmää. Tietokannan aineistona oli noin 23 000 sanomalehtiartikkelia Aamulehdestä.

Tutkimusta varten muodostettiin neljä hakemistorakennetta: taivutusmuotoinen, perusmuotoinen, ositettu perusmuotoinen ja yhdistetty taivutus- ja perusmuotoinen. Erilaisten hakutapojen vertailua varten luotiin kuusi erilaista tutkimusympäristöä. (Ks. kuvio 6.) 1) **Perinteisessä hakemisessa** tiedonhakija katkaisi sanat itse. 2) **Automaattisessa katkaisussa** perusmuotoiset hakusanat syötettiin taivutusvartaloita tuottavalle ohjelmalle ja näitä vartaloita käytettiin katkaistuina hakusanoina. 3) **Seulonnassa** taivutusvartaloilla haettiin sanojen kaikki hakemistossa esiintyvät taivutusmuodot ja näillä löytyneillä taivutusmuodoilla suoritettiin kysely. Näiden kolmen tutkimusympäristön kyselyt kohdistuivat taivutusmuotohakemistoon. 4) Haettaessa **perusmuotoja ja yhdyssanojen alkuosia** hakusanat kirjoitettiin perusmuotoisina, ja jotta saataisiin myös ne esiintymät, jossa kyseessä oleva sana on yhdyssanan alkuosana, sanan taivutusvartalot liitettiin kyselyyn katkaistuina. Tämän tutkimusympäristön kyselyt kohdistuivat perusmuotohakemistoon. 5) **Perusmuotoja ja yhdyssanojen kaikkia osia** haettaessa kyselyt kohdistuivat ositettuun perusmuotohakemistoon. Hakusanat syötettiin luonnollisesti perusmuotoisina. Sen lisäksi yhdyssanoja sisältävät dokumentit haettiin siten, että hakusanaan lisättiin manuaalisesti yhdyssanan eri osia symboloiva katkaisumerkki, koska hakemistoon oli näin merkitty, missä kohtaa yhdyssanaa kyseinen sana oli esiintynyt. 6) Kuudetta tutkimusympäristöä käytettiin **ongelmakyselyiden** suorittamiseen. Siinä haettiin ensin perusmuodoilla ja yhdyssanojen osilla, ja jos tämä ei tuottanut tulosta, haettiin taivutusmuotohakemistosta automaattista katkaisua käyttäen. Ongelmakyselyt kohdistuivat yhdistettyyn taivutus- ja perusmuotohakemistoon.

Eri tutkimusympäristöissä käytettiin vielä erilaisia kyselyitä. **Peruskyselyt** muodostettiin poimimalla hakupyynnöissä esiintyneitä sanoja kyselyyn "järkevästi" eli tiedonhakijan hyvää ammattikäytäntöä noudattaen. Näistä edelleen muodostettiin **johdoskyselyitä** laajentamalla peruskyselyt johdoskyselyiksi "mikäli hakusanalle oli keksittävissä johdoksia tai se itse oli johdos, jolle löytyi läheinen kantasana" (emt., 131). Johdosten ideoinnissa oli käytetty hyväksi muun muassa suomen kielen johdinluetteloa. **Yhdyssanakyselyssä** hakusanat katkaistiin alusta ja lopusta, jotta sanan sisältävät yhdyssanat löytyisivät. **Yhdistelmäkyselyssä** yhdistettiin edellä mainitut kyselyt eli haettiin dokumentteja, jotka sisälsivät hakusanan tai sen johdosperheen jäsenen tai sellaisen yhdyssanan, jonka osana alkuperäinen hakusana tai jokin sen johdosperheen jäsen oli. Nämä kyselyt tehtiin kaikki myös ositettuina eli mikäli hakusanojen joukossa oli yhdyssanoja, ne ositettiin. Näin kyselytyyppejä oli kaikkiaan kahdeksan. Eri tutkimusympäristöissä

haettiin vain tietyillä – niihin sopivilla – kyselyillä. Eri hakuavainten yhdistämiseen kyselyissä käytettiin erikseen JA-operaattoria ja VIRKE-läheisyysoperaattoria.



KUVIO 6. Yhteenveto Alkulan (2000) tutkimuksen hakemistoista ja tutkimusympäristöistä.

Saanti oli paras ositetussa perusmuotohakemistossa ja toiseksi paras perusmuotohakemistossa. Seuraavina olivat taivutusmuotohakemiston perinteinen hakeminen ja automaattinen katkaisu. Selvästi huonoimman saannin tuotti seulontamenetelmä taivutusmuotohakemistoon. Erot olivat pieniä, eivätkä ne olleet tilastollisesti merkitseviä, paitsi huonoimman (seulonnan) osalta.

Parhaimmat tarkkuusarvot tuotti seulontamenetelmä taivutusmuotohakemistoon. Seuraavaksi parhaat arvot tulivat perusmuotohakemistosta ja ositetusta perusmuotohakemistosta. Heikoin tarkkuus oli taivutusmuotohakemiston automaattisessa katkaisussa ja perinteisessä hakemisessa. Tarkkuusarvojen väliset erot olivat hyvin pieniä, eivätkä ne olleet tilastollisesti merkitseviä eri tutkimusympäristöjen välillä.

Koska erot eri tutkimusympäristöjen välillä olivat suhteellisen pieniä, voidaan Alkulan mukaan parhaana vaihtoehtona pitää sitä, jossa kyselyt on helpointa toteuttaa. Tiedon-

hakijalle helpoimpana tutkimusympäristönä Alkula pitää perusmuotohakemistoa, koska siinä ei hakuavainten katkaisua tarvita.

Vaikka parhaat tarkkuusarvot tuotti taivutusmuotohakemiston kyselytyyppi, niin yleisesti ottaen perusmuotohakemistoon kohdistuneet kyselyt tuottivat tarkempia tuloksia. Kun samanlainen hakulause syötettiin molempiin hakemistoihin, saatiin perusmuotohakemistosta tarkempia tuloksia. Toisaalta perusmuotohakemistosta haettaessa kannattaa Alkulan mukaan aina laajentaa kyselyä johdoksilla tai yhdyssanojen osilla, koska näin saanti paranee enemmän kuin tarkkuus laskee.

Niissä tutkimusympäristöissä, joissa käytettiin sekä perus- että johdoskyselyitä, johdoskyselyt tuottivat säännönmukaisesti peruskyselyitä paremman saannin. Tarkkuus oli sen sijaan säännönmukaisesti parempi peruskyselyissä. Saannin paremmuus johdoskyselyissä oli keskimäärin suurempi – ja myös tilastollisesti merkitsevä vähintään tasolla 0.05 – kuin tarkkuuden paremmuus peruskyselyissä.

4.3 Ohjelmat luonnollisen kielen käsittelyyn

Kuten kappaleesta 4.1 kävi ilmi, on luonnollisen kielen morfologiseen käsittelyyn kehitetty erilaisia tietokoneohjelmia. Tässä tutkimuksessa on käytetty luonnollisen kielen käsittelyssä kahta ohjelmaa, joista Fintwol perusmuotoistaa syötteet ja Finstems tuottaa syötettyjen sanojen taivutusvartalot.

FINTWOL

Fintwol on morfologinen tulkintaohjelma, joka perusmuotoistaa ohjelmalle syötetyt sananmuodot. Tässä tutkimuksessa puhutaan Fintwolista jatkossa vain Twolina, koska muiden kielten Twol-ohjelmia ei työssä ole käytetty.

Twolissa on sekä sanakirja että sääntökokoelma. Ohjelman toiminta perustuu kaksitasomalliin: leksikaaliseen ja pintatasoon. Leksikaalinen taso on ohjelman sisältämän sanakirjan sanat ja pintataso sananmuotojen todelliset esiintymät. Analysoidessaan syötteitä ohjelma tutkii vastaavatko nämä tasot toisiaan. (Alkula 2000, 88.)

Käytännössä ohjelmalle syötetään haluttu sana ja tulokseksi Twol tuottaa sanan kaikki perusmuodot ja myös morfologista tietoa sanasta (sanaluokka, luku, sijamuoto, persoona). Puutteellisesti Twol kertoo myös, onko sana johdos jostain toisesta sanasta.

Esimerkiksi sananmuodolle *hauista* Twol löytää kolme perusmuotoa, joista kaksi on monikon elatiiveja ja kolmas yksikön partitiivi.

```
Twol:          hauista
               "<hauista>"
                      "haku" N ELA PL
                      "hauki" N ELA PL
                      "hauis" N PTV SG
```

Twol pitää sanojen perusmuotoina myös verbien nominaalimuotoja.

```
Twol:          kannatettavan
               "<kannatettavan>"
                      "kannattaa" DV-TTA V REF PRES PSS
                      "kannattaa" DV-TTA PCP1 PSS POS GEN SG
                      "kannatettava" A POS GEN SG
```

Taivutusmuodon *kannatettavan* ohjelma perusmuotoista kolmeen perusmuotoon, joista kaksi ensimmäistä ovat sama sana, mutta syöte on siitä kaksi eri taivutusmuotoa. Ensimmäinen on *kannattaa*-sanan tta-johdos, verbi ja passiivin preesens. Samasta perusmuodosta syöte on myös tta-johdos, 1. partisiippi, passiivin positiivin yksikön genetiivi. Twol antaa perusmuodoksi sanan *kannatettava*, josta syöte on adjektiivi, positiivi ja yksikön genetiivi.

Adverbiaalijohdinta -sti ohjelma pitää taivutusmuotona ja perusmuotoistaa kaikki sillä johdetut sanat.

```
Twol:          kauniisti
               "<kauniisti>"
                      "kaunis" ADV POS MAN
```

Kauniisti on Twolin mukaan *kaunis*-sanan tapaa ilmaisevan adverbin positiivi.

Twol-ohjelmaa kehittää ja ylläpitää Lingsoft oy[2].

[2] Ohjelman kokeiluversio on käytettävissä Internetissä osoitteessa: http://www.lingsoft.fi/cgi-bin/fintwol

FINSTEMS

Finstems tuottaa sanan taivutusmuodot syötetystä perusmuodosta sen kirjoitusasun, lähinnä viimeisten kirjainten perusteella. Näiden lisäksi ohjelma ottaa huomioon astevaihtelun. Täydellistä sanakirjaa suomen sanoista ohjelmassa ei ole. Epäsäännöllisesti taipuvista sanoista on Finstemsissa sanakirja. (Koskenniemi 1985, 84–89.)

Käytännössä ohjelmalle syötetään haluttu sana sekä sen sanaluokka (substantiivi, adjektiivi, verbi) ja Finstems tuottaa näistä sanoista sen taivutusvartalot.

Esimerkiksi *neuvotella*-verbistä ohjelma tuottaa seuraavat vartalot

```
Finstems:     neuvottele
              neuvotteli
              neuvotell
              neuvotelt
              neuvotelk
```

Joissakin tapauksissa ohjelma tuottaa vartaloita, joita ei kielessä ole käytössä lainkaan. Tällöin on kyse tilanteesta, jossa kaksi hyvin samannäköistä sanaa taipuvat eri tavalla. Esimerkiksi sana *pakkaus*, joka muistuttaa eri tavalla taipuvaa *rakkaus*-sanaa:

```
Finstems:     pakkaus             rakkaus
              pakkauks            rakkauks
              pakkaude            rakkaude
              pakkaut             rakkaut
```

Näistä *pakkaus*-sanan vartaloista vain kaksi ensimäistä on suomen kielessä käytössä.

Finstems-ohjelmaa kehittää ja ylläpitää niin ikään Lingsoft oy[3].

[3] Ohjelman kokeiluversio on käytettävissä Internetissä osoitteessa: http://www.lingsoft.fi/cgi-bin/finstems

5 Tutkimuksen kulku

5.1 Tutkimustietokanta

Tutkimus toteutettiin TUTK-tietokannassa, joka on informaatiotutkimuksen laitoksen tutkimustietokanta. Se sisältää 53 893 suomenkielistä sanomalehtiartikkelia kokotekstinä. Mukana on sanomalehtiartikkeleita Aamulehdestä, Keskisuomalaisesta ja Kauppalehdestä vuosilta 1988–1992.

TUTK-tietokannalla on valmiina olemassa Twol-ohjelmalla perusmuotoistettu hakemisto, jossa on Twolin tunnistamat sanat perusmuotoistettuina ja tunnistamattomat taivutusmuodossa. Yhdyssanat on ositettu, joten esimerkiksi hakusana *luonto* täsmäytyy sanoihin *erämaaluonto* ja *luontoarvo.*

Taivutusmuotoinen hakemisto luotiin tätä tutkimusta varten. Siinä sanat ovat täsmälleen sellaisessa muodossa kuin ne alkuperäisissä artikkeleissa ovat. Yhdyssanoja ei siis ole ositettu.

Koska yhdyssanojen käsittely hakemistoissa on erilaista, haluttiin varmistaa, että erot tuloksissa eivät johdu siitä. Erillistä osittamatonta perusmuotohakemistoa ei tarvinnut muodostaa, koska perusmuotohakemistoon on merkitty osittamattomat sanat, joten kyselyt sai kohdistumaan vain niihin. Muuttamalla perusmuotohakemiston kyselyjä hiukan saatiin ne kohdistumaan joko kaikkiin hakemiston sanoihin tai vain osittamattomiin sanoihin. Vaikka siis erillistä osittamatonta perusmuotohakemistoa ei tarvinnut perustaa, käytetään tässä tutkimuksessa yksinkertaisuuden vuoksi ilmausta osittamaton perusmuotohakemisto ikään kuin se olisi erillinen hakemisto.

Tietokannan aineistona on lehtitekstejä, joissa esiintyy hyvin paljon erilaisia merkkejä: kirjaimia, numeroita ja erikoismerkkejä. Se, mitkä niistä ovat menneet hakemistoon tietokantaa perustettaessa eli ovat siten hakukelpoisia, on selitetty Keskustalon tutkielmassa (1994, 14–22). Hakemistoon menevät pelkkiä numeroita ja pelkkiä kirjaimia sisältävät merkkijonot siten, että Twolin tunnistamat menevät perusmuodossa toiseen hakemistoon ja tunnistamattomat alkuperäisessä taivutusmuodossa toiseen hakemistoon @-merkillä varustettuina. Merkkijonot, joissa on molempia, katkeavat aina numeron ja

kirjaimen välistä. Hakemistoon menevät siis erikseen numeromerkkijono ja kirjainmerkkijono, esimerkiksi täsmälleen merkkijonoa *airb200* voidaan hakea tarkasti vain käyttämällä läheisyysoperaattoria *#1(@airb 200)*. Erikoismerkeistä ei Keskustalon (emt.) mukaan mikään muu kuin yhdysmerkki (-) mene hakemistoon ja sekin vain, kun se esiintyy kirjainmerkkijonon keskellä tai lopussa. Ei siis koskaan kirjainmerkkijonon alussa, numero-kirjainmerkkijonon tai numeromerkkijonon yhteydessä. Erikoismerkkejä sisältäviä ilmaukset hajoavat osiin aina erikoismerkin kohdalta ja niitä pystyy tarkasti hakemaan ainoastaan läheisyysoperaattoria käyttämällä. Toki hakea voi vaikka hakulauseella *#sum(@airb 200)*, mutta silloin hakuavaimet eivät välttämättä ole vierekkäin tai molemmat edes esiinny dokumentissa.

5.2 Hakukysymykset

TUTK-tietokannalle on Sormusen lisensiaatintutkimuksessa (1993) tehty 35 hakuaihetta, joille on määritelty relevantit dokumentit. Kekäläisen (1999, 58–59) tutkimuksen mukaan näistä viidessä on keskeisimpänä käsitteenä erisnimi. Tällaisia kyselyitä ei ole mieltä laajentaa johdoskyselyiksi, kun jo pelkällä erisnimellä haettaessa saadaan kaikki relevantit dokumentit. Jätin nämä hakuaiheet pois tutkimuksesta, koska johdoslaajennus oli oleellinen osa tätä tutkimusta. Tässä tutkimuksessa on siten 30 hakuaihetta, joista kyselyt on muodostettu

Kaikille hakuaiheille on määritelty relevantit dokumentit neliportaisella relevanssiasteikolla. Relevanssi on määritelty aihesisällön näkökulmasta kuitenkin niin, että juttujen käyttötarkoitus vaikuttaa niiden informaatioarvoon. Relevanssiarviot on suorittanut neljän asiantuntijan ryhmä, joista kaksi oli freelance-toimittajia ja kaksi informaatikkoja. Arvioijat suorittivat relevanssiarviot kuvitteellisessa tilanteessa, jossa heidän piti kirjoittaa sanomalehtijuttu hakuaiheesta. (Sormunen 1993, 71–73.) Näitä alkuperäisiä relevanssiarvioita on myöhemmin täydennetty Kekäläisen osittaistäsmäyttävässä järjestelmässä tehtyjen hakujen perusteella.

Relevanssiasteikko oli seuraava:

0 jutussa ei lainkaan aiheeseen liittyvää informaatiota

1 jutussa vain viitataan aiheeseen eikä informaatiota ole enempää kuin hakukysymyksessä

2 jutussa on jossain määrin aiheeseen liittyvää informaatiota tai asia on jutussa sivuteemana

3 aihe on jutussa pääteemana ja informaatiosisältö on merkittävä.

Tässä tutkimuksessa haut toteutettiin kolmella eri relevanssitasolla. **Liberaalilla relevanssitasolla** relevanteiksi dokumenteiksi katsottiin kaikki vähänkään relevantit dokumentit eli määritelty relevanssitaso oli 1, 2 tai 3. **Normaalilla relevanssitasolla** hyväksyttiin relevanteiksi arvion 2 tai 3 saaneet dokumentit. **Tiukalla relevanssitasolla** vain relevanssiarvon 3 saaneet dokumentit hyväksyttiin relevanteiksi.

Hakukysymysten **käsitetyypit** Sormunen (1993, 38–40) jakaa yksilö- ja yleiskäsitteisiin. **Yksilökäsitteet** ovat mm. henkilön, paikan tai yritysten nimiä. **Yleiskäsitteitä** ovat konkreettiset esineet ja abstraktit käsitteet. TUTK-tietokannan hakuaiheet Sormunen on jakanut neljään käsitetyyppiin seuraavasti: (suluissa kunkin käsitetyyppiin kuuluvien hakukysymysten määrä)

1. Yleiskäsitehaut
    1.1 aihehaut (8)
    1.2 maantieteellisesti rajatut aihehaut (9)
2. Yksilökäsitehaut
    2.1 organisaatiohaut (4)
    2.2 henkilönnimihaut (9)

Hakukysymykset on lueteltu liitteessä 1. Siellä on mainittu myös kunkin kyselyn käsitetyyppi ja saantikannan koko. Eri relevanssitasoilla saantikantojen koko vaihteleekin huomattavasti. Liberaalilla tasolla relevantteja dokumentteja on kaikille kyselyille yhteensä 1 953 eli keskimäärin 35,5 kappaletta kyselyä kohti. Normaalilla relevanssitasolla niitä on 1 066 ja tiukalla tasolla vain 366 kappaletta eli keskimäärin 12,2 relevanttia dokumenttia kyselyä kohti. Tiukalla relevanssitasolla on kahdeksan kyselyä, joille on määritelty enintään viisi relevanttia dokumenttia.

5.3 Kyselyiden muodostaminen

Kyselyjä muodostettiin kolme sarjaa: ositettuun perusmuotohakemistoon, osittamattomaan perusmuotohakemistoon ja taivutusmuotohakemistoon. Kuhunkin hakemistoon tehtiin sekä perus- että johdoskysely. Kyselyistä käytetyt koodit selviävät taulukosta 1.

TAULUKKO 1. Kyselytyyppien koodit

	perusmuotohakemisto		taivutusmuotohakemisto
	ositettu	ei-ositettu	
peruskyselyt	Po1	Pe1	T1
johdoskyselyt	Po2	Pe2	T2

Silloin kun on tarve viitata sekä perus- että johdoskyselyyn samassa hakemistossa käytetään esimerkiksi ilmausta *Po-kyselyt*, joka tarkoittaa siis kyselytyyppejä *Po1 ja Po2*.

Kyselyiden muodostamisessa periaatteena oli toimia mahdollisimman loogisesti, niin että tietokonekin pystyisi samalla tavalla kyselyt muodostamaan. Näin siksi, että todennäköisesti tulevaisuuden tiedonhakujärjestelmissä kone muodostaa kyselyt itse suoraan tiedontarpeen esityksestä ilman ihmisen päättelyä.

Peruskyselyt perusmuotohakemistoon muodostettiin ottamalla lähtökohdaksi hakukysymykset, joissa siis kuvitteellinen tiedontarve on muodostettu kysymykseksi. Esimerkiksi hakukysymys numero 2:

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

Muodostamalla kyselyt suoraan kysymyksistä, ikään kuin automaattisesti, ollaan lähimpänä tilannetta, jossa tietokone itse muodostaisi kyselyt.

Hakukysymykset otettiin siinä muodossa kuin ne ovat olleet relevanssiarvioita tehtäessä. Alkuperäiset kysymykset on esitetty Sormusen (1993) lisensiaatintutkielmassa, jota varten on myös alkuperäiset relevanssiarviot tehty. Tämän jälkeen on relevanssiarvioita täydennetty, mutta hakukysymykset ovat pysyneet samana yhtä lukuun ottamatta. Teh-

tävän 26 osalta otettiin kyselyn muodostamisen lähtökohdaksi hakukysymyksen korjattu sanamuoto.

Hakukysymyksien käsittelyn ensimmäisenä vaiheena niissä esiintyneet "hajotetut" yhdyssanat yhdistettiin. Tämä tarkoitti esimerkiksi fraasin *traktori- ja kuljetusvälinetuotannon* muuttamista muotoon *traktorituotannon ja kuljetusvälinetuotannon*. Jotta Twol paremmin tunnistaisi hakukysymyksien sanoja, tehtiin kaksi pientä toimenpidettä: kaikki erisnimilyhenteet muutettiin pienaakkosiksi ja pari oikeakielisyysasiaa korjattiin. Erisnimilyhenteet piti muuttaa pienaakkosiksi, koska Twol ei tunne esimerkiksi sanaa *USA:ssa*, mutta muodon *usa:ssa* se tuntee. Oikeakielisyyteen liittyen korjattiin sana *Gorbatsov* muotoon *Gorbatshov* ja *opec:n* muotoon *opecin*. Nämäkin muutokset auttoivat Twolia tunnistamaan sanat paremmin. Näiden muutosten jälkeen sanat perusmuotoistettiin Twol-ohjelman avulla. Perusmuotoistaminen saattoi tuottaa useamman kuin yhden perusmuodon, jolloin kaikki otettiin mukaan kyselyihin.

Kaikkia sananmuotoja Twol ei tuntenut. Näitä **ongelmasanoja** hakukysymyksissä olivat:

iliescun	*2+4-neuvotteluja*
transtech	*untag-joukkojen*
bildt	*ktm:n*
bildtin	

Näistä vasemmanpuoleisen sarakkeen sanat liitettiin kyselyyn sellaisenaan @-merkillä varustettuna, koska näin TUTK-tietokannassa on merkitty tunnistamattomat sananmuodot. Oikeassa sarakkeessa olevat sanat lisättiin kyselyihin siinä muodossa kuin ne menevät hakemistoon tietokannan dokumenteissa esiintyessään. Merkkijonon *2+4-neuvotteluja* käsittelyssä erotetaan osat *2*, *4* ja *neuvotteluja*, joista *neuvotteluja*-sanan Twol perusmuotoistaa, joten hakemistoon menevät osat *2*, *4* ja *neuvottelu*. Nämä osat saavat peräkkäiset osoitteet, joten ne voidaan yhdistää uudelleen hakulausekkeella *#1(2 4 /neuvottelu)*. Lisäämällä vinoviivan *neuvottelu*-sanan eteen Inquery hakee vain sellaisia merkkijonoja, joissa *neuvottelu*-sana esiintyy yksin ilman muita sanoja. Tällöin merkkijonon *2+4-ulkoministerineuvottelu* kaltaiset ilmaisut eivät täsmäydy hakuavaimeen. Merkkijonossa *untag-joukkojen* erotetaan osat *untag* ja *joukkojen*. Näistä ensin mainittua Twol ei tunnista, joten se saa @-merkin eteensä. Sen sijaan *joukkojen*-

sanan Twol tuntee ja perusmuotoistaa. Nämä voidaan yhdistää hakulauseessa *#0(@untag /joukko)*. Vinoviivan tehtävä on tässä sama kuin edellä: estää täsmäytyminen esimerkiksi sanaan *untag-rauhanturvajoukko*. Merkkijonon *ktm:n* käsittelyssä erotetaan osat *ktm* ja *n*. Twol ei tunnista *ktm*-sanaa, joten merkkijonoa haettaessa täytyy hakulause kirjoittaa muotoon *#1(@ktm n)*. Yksittäisen kirjaimen eteen ei tule @-merkkiä.

Perusmuotoistamisen jälkeen poistettiin sulkusanat. Suomenkielinen sulkusanalista on ollut käytössä informaatiotutkimuksen laitoksen tutkijoiden CLEF-projektissa (Cross Language Evaluation Forum). Se on käännetty Inqueryn mukana tulleesta englanninkielisestä sulkusanalistasta. Sanalistaan jouduttiin tekemään muutama lisäys. Hakukysymyksissä mukana olleet yleisnimilyhenteet (*mm, jne, ym* ja *oy*) lisättiin listaan. Lisäysten jälkeen sulkusanalistassa oli 777 sanaa.

Samassa hakukysymyksessä saattoi sama sana esiintyä useampaan kertaan. Kukin sana hyväksyttiin kyselyihin mukaan vain yhden kerran. Tämän vuoksi ei otettu myöskään mukaan Twolin verbeistä tuottamia nominaalimuotoja, koska dokumentissa esiintyvä nominaalimuoto kyllä täsmäytyy kyselyn verbin perusmuotoon muutenkin.

Kaikki näin saadut sanat yhdistettiin toisiinsa #sum-operaattorilla, poikkeuksena edellä mainitut *#1(2 4 /neuvottelu)*, *#0(@untag /joukko)* ja *#1(@ktm n)*, jotka nekin yhdistettiin muiden sanojen kanssa samaan #sum-operaatioon. Hakukysymyksissä oli sanoja yhteensä 534 ja sulkusanojen poiston ja toisaalta perusmuotoistamisessa syntyvien uusien sanojen lisäyksen jälkeen kyselyssä Po1 oli sanoja 392 kappaletta eli keskimäärin 13,1 kyselyä kohti.

Osittamattoman perusmuotohakemiston kyselyt oli helppo muodostaa Po1-kyselyistä; kaikkien sanojen eteen pantiin vain vinoviiva-merkki (/). Tällöin kyselyn sanat täsmäytyivät vain hakemiston osittamattomiin sanoihin, jotka siis esiintyivät itsenäisinä sanoina myös dokumenteissa. Sanoja, joita Twol ei tunnistanut, ei pystynyt etsimään osittamattomina. Näiden kohdalla vinoviiva-merkillä ei ollut merkitystä.

Taivutusmuotohakemistoon kohdistuvien **peruskyselyiden** muotoilu oli monimutkaisempaa, sillä Inqueryssa ei toimi merkkijonokatkaisu. Katkaisua piti simuloida hake-

malla käänteistiedostosta sanojen taivutusvartaloiden avulla kaikki sanan taivutusmuodot. Tämä tapa vastaa Alkulan (2000) tutkimuksessaan käyttämää seulontahakumenetelmää. Taivutusvartalot saatiin Internetissä olevasta Finstems-ohjelmasta. Sanojen kaikkien taivutusmuotojen hakeminen käänteistiedostosta ei ollut kokonaan manuaalista. Osaltaan siinä auttoi laboratorioinsinööri Airion tekemä Unixin grep-komentoon perustuva ohjelma, joka perusmuodon ja taivutusvartaloiden avulla etsi käänteistiedostosta kaikki sanan taivutusmuodot. Ohjelma ei kuitenkaan löytänyt luotettavasti skandinaavisia kirjaimia sisältäviä sanoja, joten ne piti hakea käänteistiedostosta manuaalisesti Unixin grep-komennolla. Yleisistä sanoista tuli pisimmillään 2 500 sananmuodon lista, joista piti löytää sanan taivutusmuodot yhdyssanojen, johdosten, ylimääräisten erikoismerkkien ja samoin alkavien muiden sanojen joukosta. Manuaalisesti haetuista sanoista tarkistettiin useasti Twol-ohjelmalla, onko löytynyt taivutusmuoto juuri haetun sanan taivutusmuoto.

Alla on esimerkki grep-komennolla saadusta listasta, josta piti manuaalisesti etsiä *öljy*-sanan taivutusmuodot. Huom. | = ö ja { = ä.

```
_|ljy-yhti|st{
_|ljy-yhti|ss{
_|ljy
_|ljy-tradingissa
+_|ljyjohdoista
_|ljyjohdannaisten
_|ljyjen
*_|ljyjalosteita
```

Taivutusmuotohomonyymiasta johtuen mukaan kyselyihin on tullut myös taivutusmuotoja, jotka toisinaan ovat sanan lähinnä teoreettisia taivutusmuotoja. Esimerkiksi *suomi*-sanan taivutusmuodoista tuli mukaan *suomia*–sana, joka on erisnimen *Suomi* monikon partitiivi (teoreettinen), *suomia*-verbin perusmuoto (harvinainen) ja *suoda*-verbin III infinitiivin monikon partitiivi (varmaan todennäköisin).

Taivutusmuotohakemiston peruskyselyt (T1) muodostettiin lisäämällä Po1-kyselyihin hakuavainten kaikki tietokannassa esiintyvät taivutusmuodot. Saman sanan taivutusmuodot yhdistettiin toisiinsa #syn-operaattorilla. Ongelmasanojen kohdalla tavoitteena oli, että Inquery löytää täsmälleen samat dokumentit kuin osittamattomasta perusmuotohakemistosta. Ongelmasanoista sanat *iliescun, transtech, bildt, bildtin* lisättiin taivutusmuotohakemiston kyselyyn sellaisenaan (ilman @-merkkiä). Esiintymän *2+4*-

neuvotteluja hakulause on taivutusmuotohakemistossa muuten samanlainen, paitsi neuvottelu-sanan tilalla on #syn-operaatiossa kaikki *neuvottelu*-sanan tietokannassa esiintyvät taivutusmuodot. Ongelmasanan *untag-joukkojen* hakulause sisältää kaikki *untag-joukko*-sanaliiton taivutusmuodot. Viimeisen ongelmasanan *ktm:n* hakulause on taivutusmuotohakemistossa sama perusmuotohakemiston kyselyssä.

Johdoskyselyiden muodostamista varten piti valita, mitä johdoksia sanoista muodostetaan. Alkula (2000, 131–132) ei väitöskirjassaan ole perustellut mitenkään johdoksien valintaa. Tässä tutkimuksessa valittiin Lepäsmaan ym. (1996) erittäin produktiivisiksi nimeämät johtimet. Valinta tapahtui lähinnä siksi, että Lepäsmaan ym. kirja on lähes 20 vuotta uudempi kuin Vesikansan (1977) johdoskirja, jossa myös nimetään erittäin produktiiviset johdokset. Lepäsmaan ym. esittämistä erittäin produktiivisista johdoksista sti-johdinta ei kannattanut ottaa mukaan, sillä Twol perusmuotoistaa automaattisesti kaikki sti-johdokset. Johdoslaajennukset tehtiin taulukossa 2 olevan johdinlistan avulla. Taulukossa iso (versaali) vokaali johtimissa tarkoittaa suomen vokaalisointuun liittyvää vaihtelua A = a tai ä, O = o tai ö ja U = u tai y.

TAULUKKO 2. Johtimet, joilla johdoslaajennukset suoritettiin

kantasanan sanaluokka	johdin	johdoksen sanaluokka	esimerkki
nomini	inen	nomini	syyskuu > syyskuinen
nomini	lAinen	nomini	ranska > ranskalainen
nomini	llinen	nomini	tieto > tiedollinen
nomini	tOn	nomini	ase > aseeton
verbi	jA	nomini	yrittää > yrittäjä
verbi	minen	nomini	tuottaa > tuottaminen
verbi	ele	verbi	rakentaa > rakennella
verbi	ile	verbi	vastata > vastailla
verbi	U	verbi	johtaa > johtua
nomini tai verbi	UtU	verbi	selvitä > selviytyä
nomini	A	verbi	öljy > öljytä
nomini	ttAin	partikkeli	alue > alueittain

Kaikista kyselytyypin Po1 sanoista muodostettiin edellä mainittujen johtimien avulla uudet johdokset. Lisäksi haluttiin löytää jo kyselytyypissä Po1 esiintyneiden johdosten kantasanat, ja lisätä nämä ja kantasanan uudet johdokset johdoskyselyyn. Mutta miten määritellä, mitkä kyselytyypin Po1 sanoista ovat johdoksia? Kun mennään tarpeeksi

kauaksi kielen historiassa, useimmat sanat ovat johdoksia. Näitä ei vain enää mielletä johdoksiksi, esimerkiksi *tie* > *tietää* ja *käsi* > *käsittää*. Asian ratkaistiin muodostamalla kantasana niistä kyselytyypin Po1 sanoista, joita Twol pitää johdoksina. Vain ensimmäisen asteen kantasanat muodostettiin, ja niille edelleen kaikkien taulukossa 2 olevien johtimien avulla johdokset.

Esimerkkinä tutkimuksen johdoslaajennuksesta ovat kyselyissä esiintyneet sanat *ase* ja *puuhailla*. (Ks. Taulukko 3). *Ase*-sanasta pystyi muodostamaan kaikki muut denominaalit johdokset paitsi UtU- ja A-johdokset, koska infinitiivejä *aseutua* ja *aseta* ei ole olemassa. *Puuhailla*-verbistä pystyi muodostamaan vain kaksi deverbaalia johdosta *puuhailija* ja *puuhaileminen*. Koska Twol kertoo, että *puuhailla* on ile-johdos:

```
Twol: "puuhailla" DV-ILE V INF1 NOM,
```

sille löytyi kantasanan *puuhata*, josta muodostettiin vielä uudet johdokset. Deverbaalei-ta johdoksia syntyi *puuhaaja, puuhaaminen, puuhailla* ja *puuhautua*. *Puuhata*-sanasta ei ele- ja U-johdoksia pysty muodostamaan.

TAULUKKO 3. Sanojen ase ja puuhailla käsittely johdoslaajennuksessa

esiintynyt sana	ase	puuhailla	
sen kantasana	-	puuhata	puuhata
inen	aseinen	-	-
lAinen	aselainen	-	-
llinen	aseellinen	-	-
tOn	aseeton	-	-
jA	-	puuhailija	puuhaaja
minen	-	puuhaileminen	puuhaaminen
ele	-	-	-
ile	-	-	puuhailla
U	-	-	-
UtU	-	-	puuhautua
A	-	-	-
ttAin	aseittain	-	-
sti	aseisesti	-	-

Kantasanan etsinnässä mentiin takaisin päin vain yksi taso. Esimerkiksi

```
Twol: "yhdistyminen" DV-U DV-minen.
```

Sanan ensimmäinen kantasana on *yhdistyä*, josta muodostettiin uudet johdokset. *Yhdistyä*-sanan kantasana on *yhdistää*, mutta sitä ei otettu mukaan. Jos yhdyssanan alkuosa on Twolin tunnistama johdos, siitä ei muodostettu kantasanaa ja uusia johdoksia. Johdoksien käsittely koskee siis vain yhdyssanan viimeistä osaa. Twolin kyky tunnistaa johdoksia on varsin rajallinen. Paljon transparentteja johdoksia jäi tunnistamatta, mm. *suomalainen, tuotanto, hakemus, jäsenyys, verkosto, arvokas* jne.

Kaikista muodostetuista johdoksista karsittiin vielä harvinaisimmat pois. Karsimisen suoritin, koska näin pääsin pohtimasta kunkin johdoksen kohdalla, onko tällainen sana olemassa. Suomen kielelle on ominaista, että sanoista saa väkisin johtamalla vaikka minkälaisia johdoksia. Näistä jotkut ovat käytössä yleisesti, jotkut harvoin, ja joitakin johdoksia ei käytetä lainkaan. Karsinnan tarkoituksena oli myös vähentää ylimääräistä työtä. Karsittujen johdosten taivutusvartaloita ei tarvinnut etsiä, eikä tietenkään edelleen hakea käänteistiedostoista taivutusmuotoja. Kaikkien syntyneiden johdosten sekä löytyneiden kantasanojen esiintyminen tarkistettiin TUTKin perusmuotoisesta ositetusta hakemistosta. Jos johdos esiintyi kerrankin tietokannan hakemistoissa, se otettiin mukaan.

Kyselytyypin Po2 sanamäärä oli 846 sanaa, joten johdos- ja kantasanalaajennus lisäsi hakuavainten määrää 454:llä, joista noin 400 oli johdoksia ja loput uusia kantasanoja. Hakuavainten määrä johdoslaajennuksen jälkeen oli keskimäärin 28,2 sanaa, mikä on yli kaksinkertainen kyselytyyppiin Po1 verrattuna.

Muodostettaessa johdoskyselyä **ositettuun perusmuotohakemistoon** (Po2) lisättiin kyselytyyppiin Po1 hakuavainten kantasanat ja johdokset samaan #syn-operaatioon alkuperäisen hakuavaimen kanssa. **Osittamattoman perusmuotohakemiston** johdoskyselyt (Pe2) muodostettiin lisäämällä Po2-kyselyihin vinoviivan jokaisen hakuavaimen eteen. Johdoskyselyissä **taivutusmuotohakemistoon** (T2) samassa #syn-operaatiossa olivat alkuperäiset hakuavaimet hakemistossa esiintyneine taivutusmuotoineen sekä kantasanat ja johdokset niin ikään taivutusmuotoineen. Esimerkit hakukyselyistä ovat liitteessä 2.

5.3 Tilastollinen testaus

Jotta saaduista tuloksista pystytään sanomaan kuinka luotettavia ne ovat tai miten todennäköisesti ne johtuvat sattumasta, suoritetaan tilastollinen testaus. Se alkaa nollahypoteesin (H_0) määrittelyllä. H_0 ilmaisee, että tutkittavat menetelmät eivät eroa toisistaan. Tilastollinen testaus perustuu siihen, että tutkimuksesta saadut tulokset asetetaan nollahypoteesia vastaan. Merkitsevyystaso a määrää, millä todennäköisyydellä H_0 voidaan hylätä. Yleisiä merkitsevyystasoja ovat 0,05; 0,01 ja 0,001. Esimerkiksi merkitsevyystaso 0,01 tarkoittaa, että yhdessä tapauksessa sadasta ero johtuukin sattumasta. Kun tilastotestin tulos alittaa valitun merkitsevyystason, hyväksytään vaihtoehtoinen hypoteesi H_1. (Siegel & Castellan 1988, 7–9.) Tilastollisessa testauksessa käytetään usein tavanomaisia merkitsevyystasoja. Kun merkitsevyystaso on = 0,05, sitä sanotaan melkein merkitseväksi, = 0,01 on merkitsevä ja = 0,001 on erittäin merkitsevä.

Van Rijsbergenin (1979, 178) mukaan tiedonhaun tutkimuksessa tilastollinen testaus on ongelmallista, koska useiden suosittujen ja tehokkaiden testien ehtoja eivät tiedonhakututkimuksessa tutkittavat ilmiöt täytä. Esimerkiksi yleiset mittarit saanti ja tarkkuus eivät noudata normaalijakaumaa. Kun vertaillaan erilaisia kyselytyyppejä tai käänteistiedostoja, hakuaiheet ovat samoja eli otokset ovat toisistaan riippuvia. Tällöin, jos otoksia on kaksi, vaihtoehtoina ovat merkkitesti tai Wilcoxonin järjestyssummatesti. Näistä van Rijsbergen (1979, 179) suosittelee merkkitestiä. Jos otoksia on useampi ja ne ovat toisistaan riippuvia, käytetään usein Friedmanin testiä.

Tähän tutkimukseen on valittu tilastotestiksi Friedmanin testi. Tämä sen vuoksi, että tutkimuksessa vertaillaan 30 hakuaihetta kuudella tavalla: perus- ja johdoskyselyillä ositettuun ja osittamattomaan perusmuotohakemistoon sekä taivutusmuotohakemistoon, niinpä otoksia on kuusi ja ne ovat toisistaan riippuvia. Friedmanin testiä – ja nimenomaan sen Conoverin (1980, 299–302) esittelemää versiota – ovat käyttäneet tutkimuksissaan myös Alkula (2000) ja Kekäläinen (1999).

Friedmanin kaksisuuntainen järjestyslukutesti on merkkitestin laajennus, joka Conoverin (1980, 299) mukaan on parhaimmillaan, kun vertailtavia otoksia on viisi tai enemmän. Testissä muutetaan tutkimuksessa saatujen tulosten numeroarvot järjestysluvuiksi riveittäin siten, että huonoimmat tuloksen tuottanut menetelmä saa järjestysluvun yksi.

Sarakkeittain lasketaan järjestyslukujen summat yhteen. Tämä taulukon pohjalta Friedmanin testi selvittää, poikkeavatko vertailtavat menetelmät merkitsevästi toisistaan. (Siegel & Castellan 1988, 175–176.)

Conoverin (1980, 300) mukaan Friedmanin testin testisuureen arvo lasketaan seuraavasta kaavasta:

$$T_2 = \frac{(b-1)\left(B - bk(k+1)^2/4\right)}{A-B}, \qquad (10)$$

jossa

$$A = \sum_{i=1}^{b}\sum_{j=1}^{k}(R(X_{ij}))^2$$

ja

$$B = \frac{1}{b}\sum_{j=1}^{k}R_j{}^2$$

Kaavoissa

b = rivien (tapausten) määrä

k = sarakkeiden (menetelmien) määrä

$R(X_{ij})$ = solun järjestysluku i:nnellä rivillä, j:nnessä sarakkeessa

R_j = järjestyslukujen summa j:nnessä sarakkeessa.

Näin saatua testisuureen arvoa verrataan F-jakauman kriittiseen arvoon halutulla merkitsevyystasolla. Jos testisuureen arvo on suurempi kuin kriittinen arvo, voidaan H_0 hylätä ja edetä parittaiseen vertailuun, jossa selviää menetelmien välinen merkitsevyys. Parivertailu suoritetaan kaavalla:

$$\left|R_j - R_i\right| > t_{1-a/2}\left[\frac{2b(A-B)}{(b-1)(k-1)}\right]^{\frac{1}{2}} \qquad (11)$$

jossa

R_j ja R_i = sarakkeiden järjestyslukujen summia

$t_{1-a/2}$ = t-jakaumataulukosta saatava kriittinen arvo, merkitsevyystasolla a (oltava sama kuin aiemminkin). (Conover 1980, 299–300.)

Mellinin (1996, 162) mukaan nykyään ei merkitsevyystasoja enää yleensä määritellä etukäteen, vaan merkitsevyyslaskelmia suorittavat ohjelmat antavat suoraan p-arvon, josta johtopäätökset voi tehdä. P-arvo antaa tarkemman kuvan testin tuloksesta kuin pelkkä merkitsevyystaso.

Tulosten käytännön merkittävyyttä mittaamaan Sparck Jones (1974, 397) on käyttänyt määrättyjä prosenttitasoja. Jos kahden menetelmän vertailussa tulokset (esimerkiksi saanti- tai tarkkuusarvot) ovat tilastollisesti merkitseviä, mutta niiden ero on alle 5 prosenttiyksikköä, erolla ei ole käytännön merkittävyyttä. Jos näiden ero on 5–10 prosenttiyksikköä, sitä kutsutaan huomattavaksi. Yli 10 prosenttiyksikön ero on olennainen.

Toisaalta Keen (1992, 498) huomauttaa, että myös tulosten johdonmukainen järjestys on tärkeä havaita, vaikka ne eivät olisikaan tilastollisesti tai käytännöllisesti merkitseviä. Siis tulosten järjestys, joka johdonmukaisesti saadaan esimerkiksi erilaisilla tutkimusympäristöillä, on huomionarvoinen.

6 Tulokset

Tulososion ensimmäisissä luvuissa 6.1.1–6.1.4 esitellään tulokset koko kyselyjoukosta kolmella relevanssitasolla. Nämä tulokset on laskettu kahdella tavalla. Ensin on laskettu tarkkuus vakioiduilla saantitasoilla. Tällöin tulosten evaluointi perustuu yhtäläiseen suoritustasoon. Tämä evaluointitapa kuvaa haun tuloksellisuutta. Toiseksi on laskettu saanti ja tarkkuus dokumenttien katkaisupisteittäin. Näistä on piirretty DCV-käyriä (document cut-off value), jotka kuvaavat käyttäjän näkemää vaivaa. Katkaisupisteet ovat viiden välein välillä 1–50. Ne on valittu sen perusteella, miten tiedonhakijoiden voidaan olettaa selaavan viitteitä. Luvussa 6.2 esitellään tulokset, jotka saatiin, kun kyselyt jaettiin neljään ryhmään hakuaiheitten käsitetyypeittäin. Hakemistojen välisiä eroja selvitellään luvussa 6.3. Viimeisessä tulosluvussa 6.4 tarkastellaan kyselyjen sanamääriä hakuaiheiden käsitetyypeittäin sekä yhdyssanojen suhteellisia osuuksia.

6.1 Koko kyselyjoukon tulokset

6.1.1 Liberaali relevanssitaso

Eri kyselyiden tarkkuuskeskiarvot vakioiduilla saantitasoilla liberaalilla relevanssitasolla ovat taulukossa 4. Parhaan tarkkuuden tuotti johdoskyselytyyppi perusmuotohakemistoon, 35,1 %. Hyvin lähellä oli peruskyselytyyppi (34,3 %) samaan hakemistoon. Näiden ero oli vain 0,8 prosenttiyksikköä. Kuitenkin näiden kyselyiden välinen ero oli suurin juuri tällä relevanssitasolla. Tämä ero ei ollut tilastollisesti mitenkään merkitsevä.

Osittamattoman perusmuotohakemiston kyselyt jäivät muutaman prosenttiyksikön päähän ositetun kyselyjen tuloksista. Tämä ero oli tällä relevanssitasolla suurin. Johdoskyselytyyppi Pe2 tuotti peruskyselytyyppiä Pe1 0,6 prosenttiyksikköä paremman tuloksen (31,9 %). Vaikka näiden välinen ero oli näin pieni, se oli silti tilastollisesti melkein merkitsevä ($p = 0{,}030$). Taivutusmuotohakemiston molemmat kyselytyypit tuottivat 31,1 %:n tarkkuuskeskiarvon.

Ositetun perusmuotohakemiston tulokset olivat kolmisen prosenttiyksikköä paremmat kuin osittamattoman perusmuotohakemiston. Kyselytyyppi Po1 oli tasan 3 prosenttiyksikköä parempi kuin Pe1. Tämä ero oli tilastollisesti erittäin merkitsevä ($p < 0{,}001$).

Perusmuotohakemiston kyselytyyppien välinen suurin ero oli Po2:n ja Pe1:n välillä 3,8 prosenttiyksikköä. Tämäkin ero oli tilastollisesti erittäin merkitsevä ($p < 0,001$). Tilastollisesti merkitsevä ($p = 0,007$) ero oli myös 3,2 prosenttiyksikön ero kyselytyyppien Po2 ja Pe2 välillä. Ositetun ja osittamattoman perusmuotohakemiston kyselyjen erot olivat suurimmat juuri liberaalilla relevanssitasolla.

Ositetun perusmuotohakemiston kyselyt tuottivat 2–3 prosenttiyksikköä parempia tuloksia kuin taivutusmuotohakemiston kyselyt. Tilastollista merkitsevyyttä oli kyselytyypin Po2 tulosten erolla (3 prosenttiyksikköä) verrattuna T1:een ja T2:een. Erot olivat vastaavasti merkitsevä ($p = 0,001$) ja melkein merkitsevä ($p = 0,028$).

Osittamattoman perusmuotohakemiston tulokset olivat hyvin lähellä taivutusmuotohakemiston tuloksia. Tarkkuuskeskiarvojen erot olivat alle prosenttiyksikön. Kuitenkin kyselytyyppien Pe1 ja T2 välinen ero oli tilastollisesti merkitsevä ($p = 0,008$). Tällä liberaalilla relevanssitasolla osittamattoman perusmuotohakemiston ja taivutusmuotohakemiston väliset ero olivat kaikkein pienimpiä.

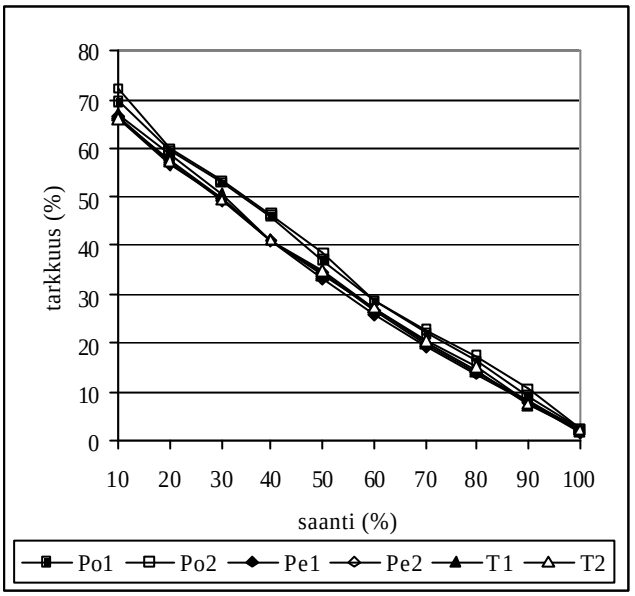
Koko liberaalin relevanssitason tilastotesti oli erittäin merkitsevä ($p < 0,001$). Mitkään liberaalin relevanssitason tilastollisesti merkitsevät ero eivät olleet Sparck-Jonesin asteikon mukaan käytännössä merkittäviä.

TAULUKKO 4. Kyselyiden tarkkuuskeskiarvot (%) sekä niiden erot (prosenttiyksikköä) vakioiduilla saantitasoilla liberaalilla relevanssitasolla

	perusmuotohakemisto				taivutusmuotohakemisto	
	ositettu		ei-ositettu			
	Po1	Po2	Pe1	Pe2	T1	T2
tarkkuuskeskiarvo	**34,3**	**35,1**	**31,3**	**31,9**	**32,1**	**32,1**
Po1	–	0,8	-3,0	-2,4	-2,2	-2,2
Po2	–	–	-3,8	-3,2	-3,0	-3,0
Pe1	–	–	–	0,6	0,8	0,8
Pe2	–	–	–	–	0,2	0,2
T1	–	–	–	–	–	0,0

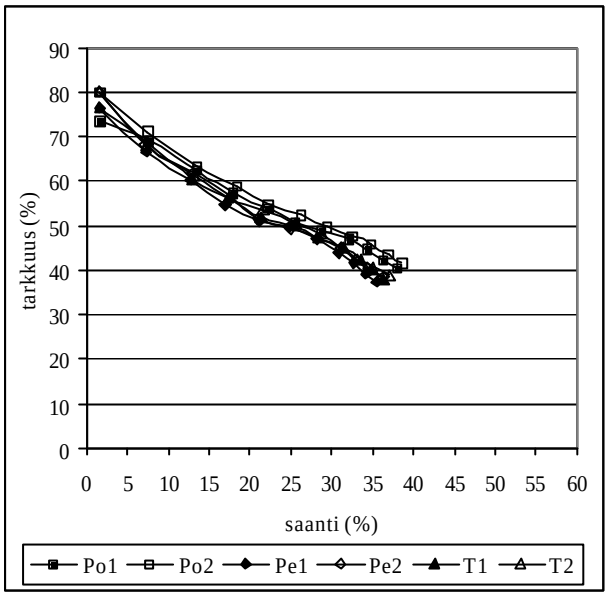
Saanti-tarkkuuskäyrässä (kuvio 7) ositetun perusmuotohakemiston kyselyjen käyrät ovat havaittavasti ylimpinä. Selvä ero tarkkuudessa muihin käyriin on erityisesti 30–

50 % saannin välillä. Osittamattoman perusmuotohakemiston ja taivutusmuotohakemiston kyselyjen käyrät ovat lähes päällekkäin. Käyrien tarkkuusarvot 10 %:n saantitasolla ovat noin 65–70 %. Käyrien toisessa päässä sadan prosentin saantitasolla tarkkuusarvot ovat kahden prosentin paikkeilla.



KUVIO 7. Saanti-tarkkuuskäyrä liberaalilla relevanssitasolla

Liberaalin relevanssitason DCV-käyrässä (kuvio 8) kaikkien kyselyjen käyrät ovat melko lähekkäin. Havaittavasti on kyselytyypin Po2 käyrä parhaimpana. Lähestyttäessä 50:tä dokumenttia myös kyselytyypin Po1 käyrä erottuu muista. Saannin osalta päästään 50:nnen dokumentin kohdalla ositetussa perusmuotohakemistossa vajaaseen 40 %:iin, jolloin tarkkuus on hiukan yli 40 %. Taivutusmuotohakemiston ja osittamattoman perusmuotohakemiston käyrien lopussa saanti on reilut 35 % ja tarkkuus hiukan alle 40 %. Tällä relevanssitasolla tarkkuusarvot olivat keskimäärin kaikkein korkeimpia ja saantiarvot matalimpia. Hakemistojen sisällä perus- ja johdoskyselyjen ero saannin osalta oli pienin kaikissa kyselytyypeissä.



KUVIO 8. DCV-käyrä liberaalilla relevanssitasolla

6.1.2 Normaali relevanssitaso

Tarkkuuskeskiarvot kyselytyypeittäin vakioiduilla saantitasoilla normaalilla relevanssitasolla ovat taulukossa 5. Ositetun perusmuotohakemiston peruskyselyjen tarkkuuskeskiarvo oli 33,0 % ja johdoskyselyjen 33,7 %. Kyselyjen ero – 0,7 prosenttiyksikköä – oli hieman pienempi kuin liberaalilla relevanssitasolla. Osittamattoman perusmuotohakemiston peruskyselyjen tarkkuuskeskiarvo oli 30,3 %, ja johdoskyselyjen 0,3 prosenttiyksikköä suurempi. Taivutusmuotohakemistossa peruskyselyjen tarkkuuskeskiarvo oli 29,7 % ja johdoskyselyjen 0,1 prosenttiyksikköä pienempi. Tämä oli ainoa kerta, kun koko kyselyjoukon kyselyissä peruskysely tuotti paremman tuloksen kuin johdoskysely.

Ositetun ja osittamattoman perusmuotohakemiston kyselyjen tarkkuuskeskiarvojen ero oli hieman pienempi kuin liberaalilla relevanssitasolla. Suurin ero (3,4 prosenttiyksikköä) oli kyselytyyppien Po2 ja Pe1 välillä. Tämä ero oli myös tilastollisesti melkein merkitsevä ($p = 0,027$).

Ositettuun perusmuotohakemistoon ja taivutusmuotohakemistoon kohdistuneiden kyselyiden erot tarkkuuskeskiarvoissa olivat kaikkein suurimpia tällä relevanssitasolla. Tarkkuuskeskiarvojen erot vaihtelivat 3,3–4,1 prosenttiyksikön välillä. Ainoa tilastollisesti merkitsevä ero oli kyselytyyppien Po2 ja T1 välillä ($p = 0,002$).

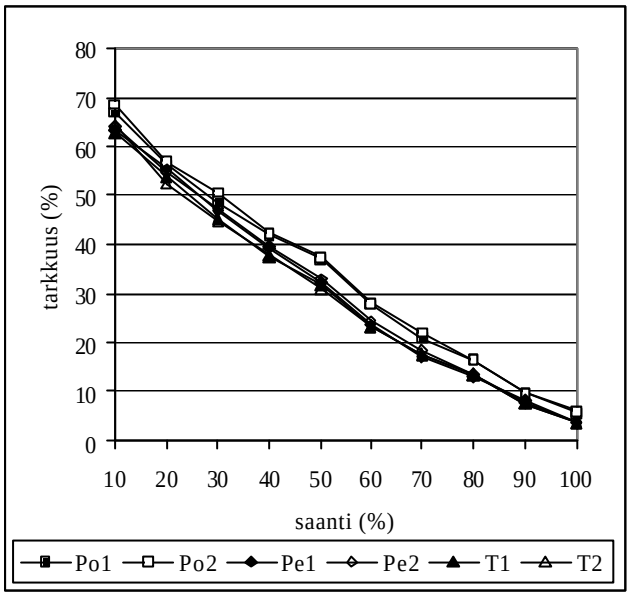
Pe-kyselyt tuottivat normaalilla relevanssitasolla aavistuksen parempia tuloksia kuin T-kyselyt. Ero oli hiukan suurempi kuin liberaalilla relevanssitasolla, mutta oli edelleen hyvin pieni; suurimmillaan vain yksi prosenttiyksikkö. Kuitenkin kyselytyyppien Pe2 ja T1 välinen ero oli tilastollisesti merkitsevä ($p = 0{,}006$).

Tämän relevanssitason tilastotesti oli melkein merkitsevä ($p = 0{,}021$). Koska tälläkin relevanssitasolla kaikki kyselyjen väliset erot jäivät alle viiden prosenttiyksikön, ei tuloksilla ole Sparck-Jonesin mukaan käytännön merkittävyyttä.

TAULUKKO 5. Kyselyiden tarkkuuskeskiarvot (%) sekä niiden erot (prosenttiyksikköä) vakioiduilla saantitasoilla normaalilla relevanssitasolla

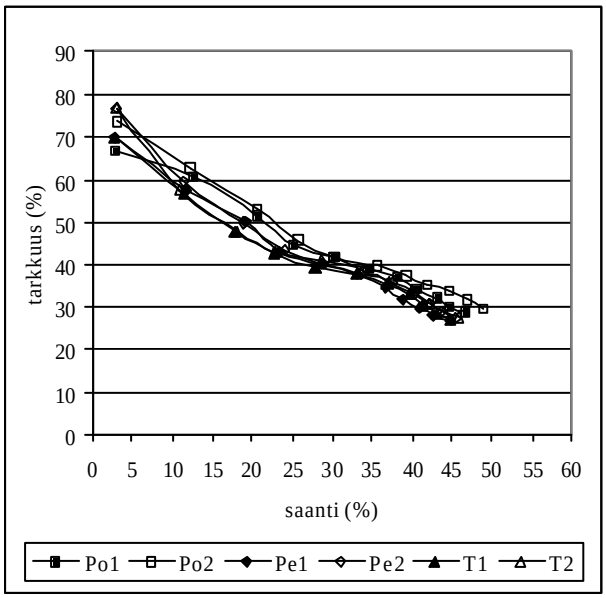
	perusmuotohakemisto				taivutusmuotohakemisto	
	ositettu		ei-ositettu			
	Po1	Po2	Pe1	Pe2	T1	T2
tarkkuuskeskiarvo	**33,0**	**33,7**	**30,3**	**30,6**	**29,7**	**29,6**
Po1	–	0,7	-2,7	-2,4	-3,3	-3,4
Po2	–	–	-3,4	-3,1	-4,0	-4,1
Pe1	–	–	–	0,3	-0,6	-0,7
Pe2	–	–	–	–	-0,9	-1,0
T1	–	–	–	–	–	-0,1

Kuviossa 9 on kyselyjen saanti-tarkkuuskäyrä normaalilla relevanssitasolla. Se on pääpiirteissään samanlainen kuin liberaalin relevanssitason käyrä. Ositetun perusmuotohakemiston käyrät kulkevat vähän muiden yläpuolella. Loput käyrät ovat käytännössä päällekkäin. Kymmenen prosentin saantitasolla kyselytyyppien tarkkuudet ovat 60 ja 70 prosentin välillä. Käyrien toisessa päässä kaikkien relevanttien dokumenttien löytymisen jälkeen tarkkuus on pudonnut noin viiteen prosenttiin.



KUVIO 9. Saanti-tarkkuuskäyrä normaalilla relevanssitasolla

Kyselyjen DCV-käyrät laskevat kuviossa 10 voimakkaasti 15:nteen dokumenttiin asti. Kyselytyypin T2 tarkkuus alenee tällä välillä lähes 34 prosenttiyksikköä.15:nnen dokumentin jälkeen käyrien lasku tasaantuu, kunnes taas 25:nnen dokumentin jälkeen lasku jälleen voimistuu. Tällä relevanssitasolla kaikkien kyselytyyppien tarkkuuskeskiarvot laskivat eniten ensimmäisestä dokumentista 50:nteen. Kyselytyypin T2 tarkkuus laski jopa 49 prosenttiyksikköä. Saannin osalta kyselyissä päästiin 50:nnen dokumentin kohdalla 45–50 %:iin. Tällöin tarkkuusarvot olivat vajaat 30 %.



KUVIO 10. DCV-käyrä normaalilla relevanssitasolla

6.1.3 Tiukka relevanssitaso

Tälläkin relevanssitasolla parhaan tuloksen – vakioiduilla saantitasoilla mitattuna – tuotti ositetun perusmuotohakemiston johdoskyselytyyppi, 23,4 %. Sen ero saman hakemiston peruskyselytyyppiin (0,6 prosenttiyksikköä) oli aavistuksen pienempi kuin normaalilla relevanssitasolla. Osittamattoman perusmuotohakemiston johdoskyselytyypin tarkkuuskeskiarvo oli 22,0 % ja peruskyselytyypin 0,4 prosenttiyksikköä huonompi. Taivutusmuotohakemiston johdoskyselytyypin tarkkuuskeskiarvo oli 21,1 %. Peruskyselytyypin saavuttama tulos oli 0,5 prosenttiyksikköä huonompi. Tällä relevanssitasolla taivutusmuotohakemiston kyselyjen välinen ero oli kaikkein suurin.

Ositetun ja osittamattoman perusmuotohakemiston kyselyjen ero oli tällä relevanssitasolla kaikkein pienin, ja oli suurimmillaan 1,8 prosenttiyksikköä. Tilastollisesti merkitseviä eroja ei näiden hakemistojen kyselyjen välillä ollut.

Hakemistojen Po ja T väliset erot olivat tällä relevanssitasolla kaikkein pienimmät. Kyselytyyppien Po1 ja T1 ero oli prosenttiyksiköissä mitattuna täysin sama kuin liberaalilla relevanssitasolla, 2,2 prosenttiyksikköä. Kuitenkin kun tarkastellaan näiden tarkkuuskeskiarvojen prosentuaalisia eroja, ne ovat suuremmat tiukalla relevanssitasolla. Liberaalilla relevanssitasolla kyselytyypin Po1 tarkkuuskeskiarvo on 6,9 % suurempi kuin kyselytyypin T1, kun se tiukalla tasolla on 10,7 % suurempi. Kyselytyyppi Po1 oli noin 2 prosenttiyksikköä parempi kuin T1 ja T2. Ero kyselytyyppiin T1 oli tilastollisesti melkein merkitsevä ($p = 0,019$). Kyselytyypin Po2 erot taivutusmuotohakemiston kyselytyyppeihin olivat kahden ja kolmen prosenttiyksikön välillä. Kyselytyypin Po2 ero T1:een oli tilastollisesti merkitsevä ($p = 0,004$).

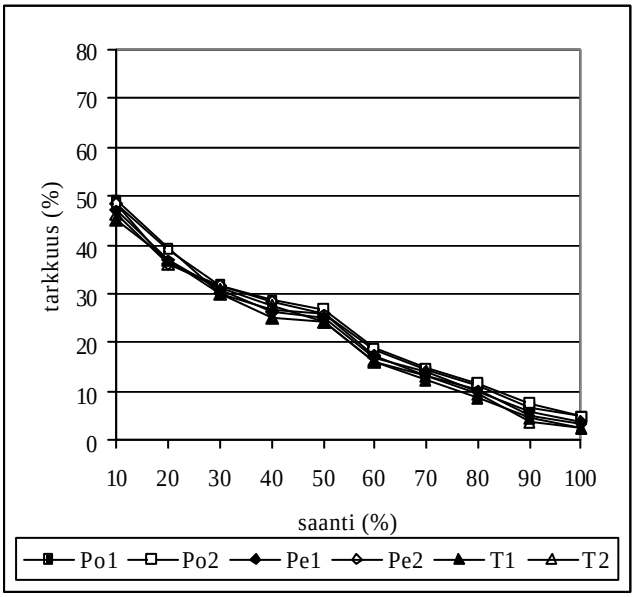
Osittamattoman perusmuotohakemiston ja taivutusmuotohakemiston kyselyjen ero oli suurin tällä relevanssitasolla. Siitä huolimatta erot eivät olleet lainkaan tilastollisesti merkitseviä. Suurimmat erot olivat kyselytyypin Pe1 ja T1 välillä (1,0 prosenttiyksikköä) sekä kyselytyyppien Pe2 ja T1 välillä (1,4 prosenttiyksikköä).

Tiukan relevanssitason tilastotestin tulos ei ollut merkitsevä ($p = 0,067$). Käytännön merkittävyyttä ei tälläkään relevanssitasolla ollut minkään kyselytyyppien välillä.

TAULUKKO 6. Kyselyiden tarkkuuskeskiarvot (%) sekä niiden erot (prosenttiyksikköä) vakioiduilla saantitasoilla tiukalla relevanssitasolla

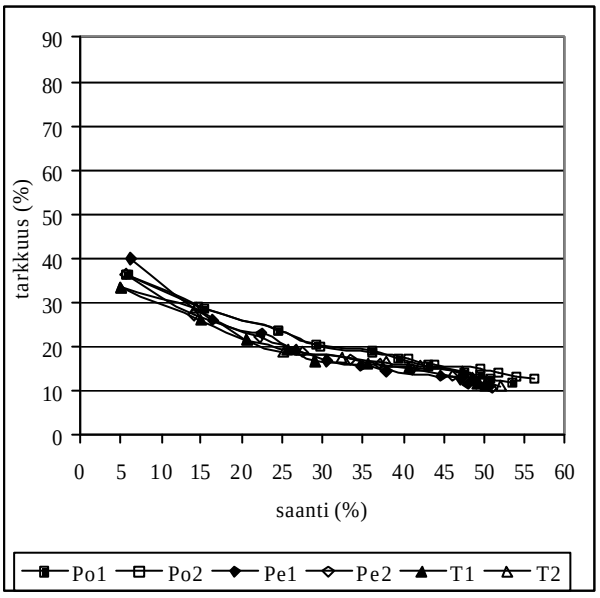
	perusmuotohakemisto				taivutusmuotohakemisto	
	ositettu		ei-ositettu			
	Po1	Po2	Pe1	Pe2	T1	T2
tarkkuuskeskiarvo	**22,8**	**23,4**	**21,6**	**22,0**	**20,6**	**21,1**
Po1	–	0,6	-1,2	-0,8	-2,2	-1,7
Po2	–	–	-1,8	-1,4	-2,8	-2,3
Pe1	–	–	–	0,4	-1,0	-0,5
Pe2	–	–	–	–	-1,4	-0,9
T1	–	–	–	–	–	0,5

Tiukalla relevanssitasolla saanti-tarkkuuskäyrä (kuvio 11) laskee kaikkein loivimmin, koska tarkkuuskeskiarvot alhaisilla saantitasoilla ovat kaikkein matalimpia. Kyselyiden tarkkuuskeskiarvon lähtötaso 10 %:n saantitasolla on vajaat 50 %. Käyrien loppupäässä kaikkien relevanttien dokumenttien löydyttyä tarkkuusarvot ovat suunnilleen samalla tasolla kuin normaalilla relevanssitasolla. Kyselyjen tarkkuuskeskiarvot alenevat hyvin hitaasti 30:n ja 50 %:n saantitason välillä, siksi saanti-tarkkuuskäyräkään ei laske tasaisesti.



KUVIO 11. Saanti-tarkkuuskäyrä tiukalla saantitasolla

Kuviossa 12 on tiukan relevanssitason DCV-käyrä. Tällä relevanssitasolla kyselyiden tarkkuuskeskiarvot olivat kaikkein alhaisimpia ja laskivat yleisesti ottaen kaikkein vähiten ensimmäisestä dokumentista 50:nteen. Esimerkiksi kyselytyypin T2 tarkkuuskeskiarvo laski vain noin 22 prosenttiyksikköä. Kaikkien hakemistojen sisällä perus- ja johdoskyselyjen saantiarvot erosivat toisistaan eniten tällä tasolla. Ositetun perusmuotohakemiston ja taivutusmuotohakemiston kyselyiden tarkkuuksien erot olivat pienimpiä tällä relevanssitasolla, toisaalta saannin osalta tämä ero oli suurin. Kyselyiden osalta päästiin saannissa noin 50–56 %:iin 50:nnen dokumentin kohdalla. Tällöin tarkkuuskeskiarvot olivat kaikissa kyselytyypeissä vähän yli 10 %:n.



KUVIO 12. DCV-käyrä tiukalla relevanssitasolla.

6.1.4 Koko kyselyjoukon tulosten yhteenvetoa

Kaikkien kyselyjen tarkkuuskeskiarvot vakioiduilla saantitasoilla olivat keskimäärin korkeimpia liberaalilla relevanssitasolla ja alhaisimpia tiukalla relevanssitasolla. DCV-käyrien tarkkuusarvot olivat parhaimmat liberaalilla relevanssitasolla ja matalimmat tiukalla tasolla. Saannin osalta tilanne oli täysin päinvastainen. Tämä on luonnollista, koska liberaalilla saantitasolla kaikki relevantit dokumentit eivät aina mahdu 50 parhaiten sijoittuneen dokumentin joukkoon. Toisaalta tiukalla relevanssitasolla saantikantojen keskimääräinen koko oli vain 12,2 dokumenttia ja siksi tarkkuus jää hyvin alhaiseksi, varsinkin loppupuolen katkaisupisteissä.

Tilastotestien mukaan merkitseviä tuloksia oli eniten liberaalilla relevanssitasolla, jossa niitä oli 7. Normaalilla tasolla niitä oli 3 ja tiukalla relevanssitasolla 2. Näistä yhteensä 12:sta merkitsevästä erosta seitsemässä toisena osapuolena oli Po2-kyselytyyppi. Kaikilla kolmella relevanssitasolla vain kyselytyyppien Po2 ja T1 välinen ero oli tilastollisesti merkitsevä. Liberaalilla ja normaalilla relevanssitasolla kyselytyyppien Po2 ja Pe1 välinen ero oli tilastollisesti vähintään melkein merkitsevä. Kaikki muut tilastollisesti merkitsevät erot esiintyivät vain yhdellä relevanssitasolla. Koko kyselyjoukon merkitsevyystestien yhteenveto on taulukossa 7.

TAULUKKO 7. Koko kyselyjoukon Friedmanin testin tulosten yhteenveto. Tilastollisesti melkein merkitsevät tulokset on lihavoitu, merkitsevät alleviivattu ja erittäin merkitsevät rasteroitu.

	relevanssitaso		
	liberaali	normaali	tiukka
testin p-arvo	**0,00009939**	**0,02101777**	0,06684973
Po1 – Po2	0,14041728	0,24294406	0,59737241
Po1 – Pe1	**0,00074530**	0,28826159	0,16020834
Po1 – Pe2	0,20969456	0,39524364	0,30781686
Po1 – T1	0,07145554	0,05261958	**0,01938415**
Po1 – T2	0,45978647	0,94345188	0,30781686
Po2 – Pe1	**0,00000423**	**0,02672744**	0,05420393
Po2 – Pe2	**0,00687110**	0,74961442	0,12265944
Po2 – T1	**0,00122595**	**0,00213981**	**0,00439614**
Po2 – T2	**0,02772826**	0,21569330	0,12265944
Pe1 – Pe2	**0,03038627**	0,05700999	0,69844490
Pe1 – T1	0,10514861	0,37589949	0,34226006
Pe1 – T2	**0,00764412**	0,32149398	0,69844490
Pe2 – T1	0,57914990	**0,00569493**	0,18200880
Pe2 – T2	0,60467398	0,35715705	1,00000000
T1 – T2	0,28429282	0,06170368	0,18200880

Kun laskettiin tarkkuuskeskiarvo läpi saantitasojen, niin kaikilla relevanssitasoilla parhaan tuloksen tuotti kyselytyyppi Po2 ja toiseksi parhaan Po1. Tällainen tendenssi on Keenin (1992) mukaan tärkeä havaita, vaikka erot olisivatkin pieniä. Kahdella relevanssitasolla kolmesta seuraavina tulivat kyselytyypit Pe2 ja Pe1. Nämä osittamattoman perusmuotohakemiston kyselytyypit olivat aina nimenomaan tässä järjestyksessä. Sen sijaan taivutusmuotohakemiston kyselytyypit eivät olleet missään johdonmukaisessa järjestyksessä.

Ositetun ja osittamattoman perusmuotohakemiston kyselyjen erot vaihtelivat 0,8–3,8 prosenttiyksikön välillä. Erot olivat suurimmat liberaalilla ja pienimmät tiukalla relevanssitasolla. Tilastollista merkitsevyyttä löytyi liberaalilla tasolla kyselytyyppien Po1–Pe1, Po2–Pe1 ja Po2–Pe2 väliltä. Näistä kaksi ensin mainittua olivat erittäin merkitseviä ja kolmas merkitsevä. Normaalilla relevanssitasolla tilastollista merkitsevyyttä (melkein merkitsevä) oli enää yhdessä vertailuparissa (Po2–Pe1) ja tiukalla tasolla ei lainkaan.

Ositetun perusmuotohakemiston ja taivutusmuotohakemiston kyselyjen erot olivat kaikkein suurimpia. Ne vaihtelivat 1,7–4,1 prosenttiyksikön välillä. Suurimmat erot olivat normaalilla relevanssitasolla. Tilastollisesti merkitseviä ero esiintyi kyselytyyppien Po1–T1, Po2–T1 ja Po2–T2 välillä. Ensin mainittu pari oli tilastollisesti melkein merkitsevä vain tiukalla relevanssitasolla. Kyselyjen Po2–T1 ero oli merkitsevä kaikilla relevanssitasoilla. Kolmannen parin ero oli tilastollisesti melkein merkitsevä liberaalilla relevanssitasolla.

Osittamattoman perusmuotohakemiston ja taivutusmuotohakemiston kyselyjen erot vaihtelivat 0,2–1,4 prosenttiyksikön välillä. Näiden hakemistojen kyselyjen tuloksellisuus meni ristiin siten, että liberaalilla relevanssitasolla taivutusmuotohakemiston kyselyt tuottivat paremmat tulokset ja kahdella muulla tasolla taas osittamaton perusmuotohakemiston kyselyt. Suurimmat erot olivat tiukalla relevanssitasolla, jossa ei kuitenkaan ollut tilastollista merkitsevyyttä. Sen sijaan merkitsevyyttä oli liberaalilla relevanssitasolla kyselytyyppien Pe1 ja T2 välillä ja normaalilla tasolla kyselytyyppien Pe2 ja T1 välillä.

Hakemistojen sisällä perus- ja johdoskyselyjen tuloksellisuudessa ei ollut juuri minkäänlaisia eroja. Suurimmillaankin erot olivat vain 0,8 prosenttiyksikköä. Kaikilla perusmuotohakemiston kyselyillä tarkkuuskeskiarvot olivat hitusen parempia peruskyselyissä, mutta yksittäisillä saantitasoilla saattoi olla toisinkin päin. Hakemistojen sisällä millään perus- ja johdoskyselyn keskinäisellä erolla ei ollut minkäänlaista tilastollista merkitsevyyttä.

Kaikki kyselyjenväliset erot tarkkuuskeskiarvojen välillä jäivät alle viiden prosenttiyksikön eikä näillä tuloksissa ole siten Sparck-Jonesin mukaan käytännön merkittävyyttä.

Koska osittaistäsmäyttävissä tiedonhakujärjestelmissä ei toimi hakuavainten katkaisu, piti tässä tutkimuksessa etsittäessä sanojen taivutusmuotoja hakemistosta simuloida katkaisua. Katkaisua simuloitiin hakemalla taivutusvartaloiden avulla niihin täsmäävät sananmuodot, jotka vielä joko automaattisesti perusmuotoistettiin tai käytiin manuaalisesti läpi. Näin löydettiin halutun sanan taivutusmuotohakemistossa esiintyneet sananmuodot. Tässä työssä katkaisun simulointi ei ollut täydellistä, koska aidossa katkaisussa mukaan hakuun tulevat kaikki samalla merkkijonolla alkavat sananmuodot, ja tässä sillä etsittiin vain sanan taivutusmuodot. Mekaanisena toimenpiteenä katkaisun simulointi onnistui hyvin. Taivutusmuotohakemiston kyselyjen tulokset olivat lähes samat kuin tulokset osittamattomasta perusmuotohakemistosta.

6.2 Tulokset hakuaiheiden käsitetyypeittäin

Sormunen (1993) on jaotellut TUTK-tietokannan hakuaiheet neljään käsitetyyppiin. Tässä tutkimuksessa selvitettiin, minkä käsitetyypin kyselyt ovat tuloksellisimpia, kun kyselyt muodostetaan suoraan hakukysymyksistä. Lisäksi selvitettiin, miten johdoslaajennus vaikuttaa eri käsitetyyppeihin ja miten tulokset eroavat tutkittavissa hakemistossa. Tarkkuus laskettiin vakioiduilla saantitasoilla vain normaalilla relevanssitasolla, koska edellisten tulosten perusteella voitiin olettaa, että sen antamat tulokset ovat riittäviä päätelmien tekoon. Eri kyselyitten sanamäärät ja yhdyssanojen osuudet laskettiin käsitetyypeittäin.

Kaikissa hakemistoissa parhaat tulokset tulivat henkilökyselyistä. Pe- ja T-kyselyillä organisaatiokyselyt tuottivat toiseksi parhaan tuloksen, kun taas Po-kyselyillä sen tuottivat maantieteellisesti rajatut kyselyt. Kaikkien hakemistojen huonoimmat tulokset tulivat aihekyselyistä. Tämä järjestys onkin osin pääteltävissä, koska henkilöaiheissa kohtuullinen tulos voi tulla jo sillä, että hakee henkilön nimellä. Myös maantieteellisesti rajatun aiheen kyselyissä ja organisaatioaiheisissa kyselyissä on erisnimi, joka on tärkeä hakuavain.

Kyselyjen tarkkuuskeskiarvot hakuaiheiden käsitetyypeittäin ovat taulukossa 8. **Aihekyselyitä** oli yhteensä kahdeksan. Kyselytyyppien tarkkuuskeskiarvot olivat yleisesti heikoimpia juuri aihekyselyissä. Parhaan tuloksen (23,3 %) tuotti kyselytyyppi Po2.

Aihekyselyissä erot ositetun perusmuotohakemiston ja taivutusmuotohakemiston tai osittamattoman perusmuotohakemiston välillä olivat noin 4prosenttiyksikön luokkaa. Tilastollisesti merkitseviä eroja ei ollut minkään kyselytyyppien välillä. Osittamattoman perusmuotohakemiston kyselyiden tulokset olivat lähes samat kuin taivutusmuotohakemiston kyselyjen. Koko testin tulos ei ollut tilastollisesti merkitsevä (p = 0,361).

TAULUKKO 8. Kyselyiden tarkkuuskeskiarvot (%) hakuaiheiden käsitetyypeittäin sekä niiden erot (prosenttiyksikköä) normaalilla relevanssitasolla

	perusmuotohakemisto				taivutusmuotohakemisto	
	ositettu		ei-ositettu			
	Po1	Po2	Pe1	Pe2	T1	T2
Aihekyselyt (N = 8)						
tarkkuuskeskiarvo	**22,8**	**23,3**	**19,0**	**18,9**	**19,1**	**18,8**
Po1	–	0,5	-3,8	-3,9	-3,7	-4,0
Po2	–	–	-4,3	-4,4	-4,2	-4,5
Pe1	–	–	–	-0,1	0,1	-0,2
Pe2	–	–	–	–	0,2	-0,1
T1	–	–	–	–	–	-0,3
Maantieteellisesti rajatut kyselyt (N = 9)						
tarkkuuskeskiarvo	**36,6**	**37,9**	**29,6**	**30,6**	**29,3**	**30,2**
Po1	–	1,3	-7,0	-6,0	-7,3	-6,4
Po2	–	–	-8,3	-7,3	-8,6	-7,7
Pe1	–	–	–	1,0	-0,3	0,6
Pe2	–	–	–	–	-1,3	-0,4
T1	–	–	–	–	–	0,9
Henkilökyselyt (N = 4)						
tarkkuuskeskiarvo	**44,7**	**44,0**	**45,2**	**44,8**	**40,9**	**37,0**
Po1	–	-0,7	0,5	0,1	-3,8	-7,7
Po2	–	–	1,2	0,8	-3,1	-7,0
Pe1	–	–	–	-0,4	-4,3	-8,2
Pe2	–	–	–	–	-3,9	-7,8
T1	–	–	–	–	–	-3,9
Organisaatiokyselyt (N = 9)						
tarkkuuskeskiarvo	**33,4**	**34,3**	**34,3**	**34,8**	**34,4**	**35,3**
Po1	–	0,9	0,9	1,4	1,0	1,9
Po2	–	–	0,0	0,5	0,1	1,0
Pe1	–	–	–	0,5	0,1	1,0
Pe2	–	–	–	–	-0,4	0,5
T1	–	–	–	–	–	0,9

Maantieteellisesti rajatuissa kyselyissä paras yksittäinen kyselytyyppi oli Po2, joka tuotti 37,9 % tarkkuuskeskiarvon. Ositetun perusmuotohakemiston kyselyjen tulosten paremmuus verrattuna Pe- ja T-kyselyjen tuloksiin oli suurempaa kuin muissa aiheissa. Kyselyjen väliset erot olivat suurimmillaan 8,6 prosenttiyksikköä. Kyselytyypin Po1 noin 7 prosenttiyksikön ero kyselytyyppeihin Pe1 ja T1 oli tilastollisesti melkein merkitsevä ($p = 0,048$ ja $p = 0,034$). Kyselytyypin Po2 yli 8 prosenttiyksikön ero Pe1:een ja T1:een oli tilastollisesti erittäin merkitsevä ($p < 0,001$). Lisäksi sen ero kyselytyyppi T2:een oli melkein merkitsevä ($p = 0,012$). Tilastotestin tulos oli tilastollisesti merkitsevä ($p = 0,005$). Maantieteellisesti rajattuja kyselyitä oli yhteensä yhdeksän.

Henkilökyselyitä oli vain neljä kappaletta eikä testi ollut mitenkään tilastollisesti merkitsevä. Juuri henkilökyselyissä saatiin parhaimmat tarkkuuskeskiarvot. Paras tulos oli 45,2 %, jonka tuotti kyselytyyppi Pe1. Po- ja Pe-kyselyiden välillä erot olivat hyvin pieniä. Taivutusmuotohakemiston kyselyt tuottivat puolestaan erilaiset tulokset. Peruskysely oli 3,9 prosenttiyksikköä parempi kuin johdoskysely. Näin suurta eroa ei hakemiston sisäisillä kyselyillä ollut missään muissa aiheissa eikä koko kyselyjoukon kyselyissä. Koska näin isoa eroa perus- ja johdoskyselyn välillä ei ollut kummassakaan perusmuotohakemistossa, taivutusmuotohakemiston kyselyjen ero ei voi johtua siitä, että henkilökyselyissä peruskysely olisi yleisesti parempi.

Mistä ero sitten johtuu? Henkilökyselyitä on vain neljä, joten yhdenkin peruskyselyn selvä paremmuus saattaa vaikuttaa keskiarvoon. Näin olikin. Henkilökyselyihin kuuluva kysely 33 tuotti peruskyselynä 9,3 prosenttiyksikköä paremman tarkkuuskeskiarvon kuin johdoskysely, ja kun vielä kyselyssä 34 peruskyselyn ero johdoskyselyyn oli 3,9 prosenttiyksikköä, oli kaikkien neljän kyselyn tarkkuuskeskiarvo selvästi parempi peruskyselyillä. Kävin vielä läpi kyselyn 33 selvittääkseni, mistä suuri ero muodostuu. Hain ensin peruskyselyn jokaisella sanalla erikseen. Koska kyselyt kohdistuivat taivutusmuotohakemistoon, jokainen "sana" tarkoittaa tässä sanan kaikkia hakemistossa esiintyviä taivutusmuotoja. Sitten hain johdoskyselyn sanoilla ja niiden mahdollisesti saamilla johdoksilla. Näissä hauissa ei perus- ja johdoskyselyn sanoilla ollut mitään olennaisia eroja. Sitten lisäsin kummankin kyselyn hakulauseeseen yhden sanan kerrallaan ja seurasin tarkkuuskeskiarvon kehittymistä. Suurimmat muutokset perus- ja johdoskyselyn välillä tulivat sanojen *mielipide* ja *liittyä* lisäämisen jälkeen. Sana *mielipide*

suurensi kyselyjen eroa 5,3 prosenttiyksikköä, vaikka sanalla ei ole mitään johdoksia eli molempiin kyselyihin lisättiin täsmälleen samat *mielipide*-sanan taivutusmuodot. Sana *liittyä*, joka oli johdoslaajennuksessa saanut johdokset *liittyjä* ja *liittyminen*, kasvatti kyselyjen eroa 4,6 prosenttiyksikköä. Tämä selvitys ei tuonut mitään selvää ratkaisua henkilökyselyjen perus- ja johdoskyselyjen suurehkoon eroon. Ilmeisesti vain johdoslaajennuksessa lisätyt johdokset yhdessä huononsivat hakutulosta.

Organisaatiokyselyitä oli yhteensä yhdeksän. Parhaimmat tuloksen (35,3 %) tuotti kyselytyyppi T2. Tämä oli ainoa kyselysarja aihekyselyistä ja koko kyselyjoukon kyselyistä, jossa taivutusmuotohakemiston kyselyt tuottivat parempia tuloksia kuin ositetun perusmuotohakemiston kyselyt. Kaikkien kyselyiden erot olivat pienimpiä juuri organisaatiokyselyissä. Suurimmillaankin ero oli vain 1,9 prosenttiyksikköä. Siitä huolimatta tilastollisesti merkitseviä eroja syntyi. Tämä on mahdollista, koska Friedmanin merkitsevyystestissä muutetaan tutkimuksessa saatujen tulosten numeroarvot (eli tarkkuuskeskiarvot) järjestysluvuiksi (tässä 1–6). Näin muutama hyvin pienikin ero jonkun kyselytyypin eduksi voi saada tuloksen näyttämään merkitsevältä. Kyselytyyppi Pe2:n paremmuus Po1:een verrattuna oli tilastollisesti merkitsevä ($p = 0,005$) sekä Po2:een ja T1:een verrattuna melkein merkitsevä ($p = 0,039$ ja $p = 0,010$). Organisaatiokyselyissä parhaan tuloksen tehneen kyselytyypin T2 paremmuus oli tilastollisesti merkitsevä verrattuna kyselytyyppiin Po1 ($p = 0,005$) sekä melkein merkitsevä verrattuna Po2:een ja T1:een ($p = 0,039$ ja $p = 0,010$). Organisaatiokyselyiden tilastotestin tulos oli melkein merkitsevä ($p = 0,011$). Tilastotestien yhteenveto hakuaiheiden käsitetyypeittäin on taulukossa 9.

Organisaatiokyselyissä kyselytyyppien erot olivat pieniä, koska näiden joukossa oli useampi kysely, jossa yhdyssanojen osittaminen huononsi tulosta. Kävin kaksi organisaatiokyselyihin kuuluvaa kyselyä läpi (numerot 6 ja 22), joissa erot Po- ja T-kyselyiden välillä olivat suurimmat, samalla tavalla kuin aiemmin henkilökyselyiden kohdalla on selitetty. Tämä tapa ei tosin ole kovin luotettava, sillä järjestys, jossa sanoja lisää saattaa vaikuttaa selvityksen lopputulokseen. Jotain siitä pystyy kuitenkin päättelemään. Vaikuttaa siltä, että useimmiten osituksen tulosta huonontavaan vaikutukseen ei ole yksittäistä syytä; on vain sanoja, jotka osittamattomina yhdessä tuottavat paremman tuloksen. On kuitenkin myös joitain yksittäisiä sanoja, joiden vaikutus koko kyselyn tulokseen on suuri. Tällainen on esimerkiksi kyselyn 6 *liitto*-sana. Yksistään tällä sanalla

haettuna kyselyn tarkkuuskeskiarvo on perusmuotohakemistossa 6,2 % ja taivutusmuotohakemistossa 23,8 %. Tällaisessa tapauksessa syykin on ilmeinen: koska kyselyn 6 tärkeä fraasi *Varsovan liitto* kirjoitetaan erilleen, suosii se tietenkin osittamatonta hakemistoa. Lehtitekstissä yleiset ammattiliitto-jutut tulevat helposti kärkeen ositetusta hakemistosta haettaessa. Koska näissä aihekyselysarjoissa oli kyselyitä vähän, neljästä yhdeksään, vaikuttaa jo muutaman kyselyn tulos koko kyselysarjan tuloksen keskiarvoon.

Hakemistojen sisällä perus- ja johdoskyselyitten väliset erot olivat suurimmillaan 1,3 prosenttiyksikköä (pois lukien henkilökyselyiden taivutusmuotohakemiston kyselyjen suurempi ero) eikä tilastollisesti merkitseviä eroja ollut. Myöskään mitään tendenssiä jommankumman kyselyn paremmuudesta ei ollut havaittavissa.

Taulukko 9. Friedmanin testin tulosten sekä Sparck Jonesin käytännön merkitsevyyden yhteenveto hakuaiheiden käsitetyypeittäin. Tilastollisesti melkein merkitsevät tulokset on lihavoitu, merkitsevät alleviivattu ja erittäin merkitsevät rasteroitu.

Käytännön merkitsevyys on merkitty: H = huomattava ja O = olennainen.

	Aihekyselyt	Maantiet. rajatut kyselyt	Henkilökyselyt	Organisaatiokyselyt
testin p-arvo	0,36143053	**0,00481033**	0,94749588	**0,01055082**
Po1 – Po2	1,00000000	0,11722326	0,93306506	0,39837474
Po1 – Pe1	0,23339981	**0,04822576** H	0,86664432	0,32526702
Po1 – Pe2	0,23339981	0,77247149	0,73734999	**0,00478506**
Po1 – T1	0,11489761	**0,03493989** H	0,67539525	0,77745318
Po1 – T2	0,11489761	0,31438166	0,61577511	**0,00478506**
Po2 – Pe1	0,23339981	**0,00077718** H	0,80124652	0,88757706
Po2 – Pe2	0,23339981	0,06571710	0,67539525	**0,03901297**
Po2 – T1	0,11489761	**0,00050789** H	0,73734999	0,57246387
Po2 – T2	0,11489761	**0,01236868** H	0,67539525	**0,03901297**
Pe1 – Pe2	1,00000000	0,08837271	0,86664432	0,05324292
Pe1 – T1	0,68852156	0,88500375	0,55882680	0,48097438
Pe1 – T2	0,68852156	0,31438166	0,50482690	0,05324292
Pe2 – T1	0,68852156	0,06571710	0,45398837	**0,01003212**
Pe2 – T2	0,68852156	0,47099066	0,40646011	1,00000000
T1 – T2	1,00000000	0,25114489	0,93306506	**0,01003212**

6.3 Hakemistojen erot

Selvin ja ennalta tiedetty ero hakemistojen välillä oli ositetun perusmuotohakemiston yhdyssanojen ositus. Sen sijaan osittamattoman perusmuotohakemiston ja taivutusmuotohakemiston kyselyiden pitäisi tuottaa täsmälleen samat dokumentit. Siten esimerkiksi sanan */luonto* syöttäminen osittamattoman perusmuotohakemistoon pitäisi tuottaa samat dokumentit kuin taivutusmuotohakemistoon synonyymisina syötettävät kaikki sen sanan taivutusmuodot *#syn(luonto luonnon luontoa luonnoksi jne.)*. Tuloksia tarkasteltaessa nämä hakemistot tuottivatkin lähes samanlaisia tuloksia, mutta pieniä erojakin oli.

Erot osittamattoman perusmuotohakemiston ja taivutusmuotohakemiston välillä johtuivat kahdesta asiasta: yhdysmerkin sisältävien sanojen erilaisesta käsittelystä ja virheistä etsittäessä hakemistossa esiintyviä sanojen taivutusmuotoja. Näistä ensin mainittu oli merkittävä ja aiheutti suurimman osan eroista. Jälkimmäinen syy oli hyvin harvinainen eikä aiheuttanut juuri eroja tuloksissa.

Yhdysmerkin sisältäviä yhdyssanoja osittamaton perusmuotohakemisto ja taivutusmuotohakemisto käsittelevät siis eri tavoilla. Vaikka osittamattoman perusmuotohakemiston sanojen pitäisi olla täysin alkuperäisessä asussaan, siellä oli kuitenkin ositettu yhdysmerkin sisältämät yhdyssanat. Täsmäytyksessä tämä ilmeni siten, että osittamattomassa perusmuotohakemistossa esimerkiksi sana *ukko* täsmäytyisi sanoihin *tikku-ukko* tai *ukko-parka.* Taivutusmuotohakemistossa näin ei tapahtuisi.

Tarkasteltaessa näiden hakemistojen eroja yksittäisissä kyselyissä havaittiin suurimmat erot kyselyjen 33 ja 34 kohdalla. Ositetussa perusmuotohakemistossa kyselyyn 33 täsmäävissä dokumenteissa sanat *esko* ja *aho* esiintyvät (myös) vain yhdysmerkin kanssa. Kyselyn *aho*-sana täsmäytyy *korkia-aho*-sanaan ja *ahon-kullbergin*-sanaan. Ja toisaalta *esko*-sana täsmäytyy *esko-ship*-sanaan. Kyselyssä 34 *kauko*-sana täsmäytyy melko yleiseen *kauko-itä*-sanaan. Nämä kaikki ovat tapauksia, joissa täsmäytystä ei tapahdu taivutusmuotohakemistossa.

Toisaalta myös taivutusmuotohakemiston kyselyt täsmäytyvät sanoihin, joihin osittamattoman perusmuotohakemiston kyselyt eivät täsmäydy. Tämäkin liittyy yhdysmerkin

esiintymiseen yhdyssanassa. Tällaisia sanoja, joita osittamaton perusmuotohakemiston kysely ei löydä, mutta taivutusmuotohakemiston kysely löytää, ovat mm. *hinta- [ja palkkasulkua]*, *öljy- [ja kaasujohdot]*. Siis kun yhdysmerkki esiintyy sanan perässä, hakemistot löytävät sanat eri tavalla.

Koska osittamattoman perusmuotohakemiston kyselyt tuottivat hiukan parempia tuloksia kuin taivutusmuotohakemiston kyselyt, ovat *korkia-aho*-tyyppiset esiintymät merkityksellisempiä hakujen tuloksellisuuden kannalta kuin *"hinta-"*-tyyppiset esiintymät. Ositetussa perusmuotohakemiston kyselyissä täsmäytyvät kaikki edellä mainitut tapaukset.

Taivutusmuotohakemiston kyselyiden yhdysmerkin sisältävät yhdyssanat hajoavat kyselyä suoritettaessa erilleen. Kun kyselyssä ovat hakusanan *etelä-amerikka* kaikki taivutusmuodot,

#syn(etelä-amerikka etelä-amerikan etelä-amerikassa jne.)

täsmäytyy kysely dokumentteihin, joissa on joko koko *etelä-amerikka*-sana tai sana *etelä* tai *amerikka*. Tulosjoukon koko oli 801 dokumenttia normaalilla relevanssitasolla. Täsmäytyminen ei kuitenkaan vastannut kyselyä

#sum(#syn(etelä etelän etelässä jne) #syn(amerikka amerikan amerikassa jne.)),

joka tuotti 1 379 dokumenttia. Kokonaisia *etelä-amerikka*-sanoja esiintyi 171 dokumentissa. Tällaista yhdysmerkin sisältävien yhdyssanojen hajoamista ei tapahtunut perusmuotohakemistoista haettaessa, vaan siellä sanan piti esiintyä kokonaisena dokumentissa. Perusmuotohakemistoissa siis kyselyt *etel{-amerikka ja #0(etel{ amerikka)* tuottivat saman tuloksen. Hakutehtävästä riippuen tämä ero saattoi parantaa tai huonontaa taivutusmuotokyselyn tuloksia verrattuna perusmuotohakemiston kyselyihin. Yhdysmerkin vaikutuksista eri hakemistoissa on yhteenveto taulukossa 10.

TAULUKKO 10. Yhteenveto yhdysmerkin vaikutuksesta kyselyn ja dokumentin täsmäämiseen eri hakemistoissa yhdysmerkin esiintyessä sanan eri kohdissa kyselyissä tai dokumenteissa. Rasteritausta ilmaisee odotuksenvastaisen tuloksen.

	Ositettu perusmuoto-hakemisto	Osittamaton perusmuoto-hakemisto	Taivutusmuoto-hakemisto
kysely: *hinta*			
dokumentti: *hinta- [ja palkkasulku]*	täsmää	ei täsmää	täsmää
kysely: *aho*			
dokumentti: *korkia-aho*	täsmää	täsmää	ei täsmää
kysely: *etelä-amerikka*			
dokumentti: *etelä-amerikka*	täsmää	täsmää	täsmää
dokumentti: *etelä*	ei täsmää	ei täsmää	täsmää osittain
dokumentti: *amerikka*	ei täsmää	ei täsmää	täsmää osittain

Taivutusmuotohakemiston kyselyjen muodostamisessa on varmasti tapahtunut joitain virheitä, koska hakusanojen poiminta perustui inhimilliseen päättelyyn. Hakemistossa esiintyviä taivutusmuotoja on voinut jäädä poimimatta. Tarkistuksia suoritettaessa löytyi yksi vahingossa pois jäänyt sananmuoto *öljyistä*. Todennäköisesti sananmuoto on ajateltu *öljyinen*-sanan yksikön partitiiviksi, jota ei olisikaan pitänyt ottaa mukaan, eikä ole huomattu, että se on myös *öljy*-sanan monikon elatiivi. Tällaiset virhetapaukset ovat kuitenkin vähäisiä ja satunnaisia, eikä niillä voi olla yleistä merkitystä tuloksiin.

Perus- ja taivutusmuotohakemiston kyselyissä on pientä eroa, joka johtuu siitä, että taivutusmuotohakemisto tehtiin seulomalla. Perusmuotokyselyn sanan täytyi siis esiintyä tietokannassa, jotta se tuli mukaan taivutusmuotohakemiston kyselyyn. Aina näin ei ollut. Esimerkiksi perusmuotohakemistojen kyselyssä 22 on sanat *maakaasutoiminto* ja *maakaasutoiminta*. Nämä sanat eivät kuitenkaan esiintyneet tietokannan hakemistossa, joten niiden taivutusmuotoja ei voitu liittää taivutusmuotohakemiston kyselyihin. Testien mukaan tällaiset tapaukset eivät huonontaneet perusmuotohakemistojen kyselyjen tuloksia, sillä nollatuloksen tuottava hakuavain ei vaikuttanut tulokseen mitenkään. Tulos on täysin sama kuin ilman ko. sanoja.

Tavoite siitä, että kyselyt perusmuotohakemistoihin ja taivutusmuotohakemistoon olisivat olleet yhdyssanojen ositusta lukuun ottamatta samanlaisia, ei aivan toteutunut. Tämä

johtui siitä, että hakulauseita suunniteltaessa ja testattaessa eri hakemistoihin, haut tehtiin interaktiivinen Inquery -ohjelmalla, jonka avulla tietokantaan pystyy tekemään hakuja ilman relevanssiarvioita. Varsinaiset haut tehtiin kuitenkin muiden ohjelmien avulla, jotka sitten käsittelivät yhdysmerkkiä eri tavalla. Yhdysmerkin erilainen käyttäytyminen hakemistoissa esti kyselyjen samanlaisen toiminnan hakemistoissa. Kuitenkaan sen vaikutus ei ollut koko kyselyjoukossa niin merkittävää, että se estäisi johtopäätöksien tekemisen.

6.4 Kyselyiden sanamäärät ja yhdyssanojen osuudet

Kyselyiden sanamäärät on saatu laskemalla yhteen kaikki perusmuotohakemistoihin kohdistuneen 30 kyselyn sanat kyselytyypeistä Po1 ja Po2 (käytännössä samat kuin Pe1 ja Pe2). Taivutusmuotohakemistoon kohdistuneiden kyselyiden sanamääriä ei ole mieltä laskea, koska sanamäärä riippuu aivan siitä, mitä sananmuotoja kustakin sanasta tietokannan hakemistossa esiintyy. Perus- ja johdoskyselyiden sanamäärät hakuaiheiden käsitetyypeittäin ovat taulukossa 11.

Johdoslaajennus kasvatti kyselyjen sanamäärää selvästi vähiten henkilökyselyissä, joissa sanamäärä ei edes kaksinkertaistunut. Muissa kyselyissä sanojen määrän kasvu oli suunnilleen samaa luokkaa. Organisaatiokyselyissä kasvu oli suurinta, 128 %.

Yhdyssanojen määrä laskettiin niin ikään perusmuotohakemistojen kyselyistä hakuaiheiden käsitetyypeittäin eroteltuna. Yhdyssanoja oli suhteessa eniten aihekyselyissä, 39 % kaikista kyselyn sanoista. Selvästi vähiten niitä oli henkilökyselyissä. Johdoskyselyissä yhdyssanojen suhteellinen osuus laski paljon, koska yhdyssanoilla on harvoin johdoksia. Suhteessa eniten niitä oli edelleen aihekyselyissä, mutta vähiten nyt organisaatiokyselyissä.

Taulukko 11. Perus- ja johdoskyselyiden sanamäärät yhteensä, niiden prosentuaalinen lisäys sekä yhdyssanojen suhteellinen osuus kyselyiden sanoista hakuaiheiden käsitetyypeittäin

	Aihe-kyselyt N = 8	Maantiet. rajatut kyselyt N = 9	Henkilökyselyt N = 4	Organisaatiokyselyt N = 9	yhteensä N = 30
peruskyselyjen sanamäärä	104	137	46	105	392
johdoskyselyjen sanamäärä	226	296	83	241	846
lisäys	117 %	116 %	80 %	130 %	116 %
yhdyssanoja peruskyselyjen sanoista	39 %	31 %	26 %	30 %	32 %
yhdyssanoja johdoskyselyjen sanoista	21 %	19 %	16 %	14 %	18 %

Yhdyssanojen suhteellinen määrä kyselyissä ei näytä vaikuttavan hakujen tuloksiin huolimatta perusmuotohakemistojen erilaisesta yhdyssanakäsittelystä. Hakuaiheiden käsitetyypeittäin tarkasteltuna suurin ero Po- ja Pe-kyselyiden välillä oli maantieteellisesti rajatuissa kysymyksissä. Näissä kysymyksissä yhdyssanoja oli aivan keskiarvoa vastaava määrä. Henkilö- ja organisaatiokyselyissä yhdyssanoja oli keskiarvoa vähemmän, ja näiden kohdalla perusmuotohakemistojen kyselyiden välinen ero oli hyvin pieni.

7 Keskustelua

Tämän tutkimuksen päätutkimusongelma oli, miten tiedonhaun tuloksellisuus eroaa käytettäessä perusmuotoista ja taivutusmuotoista hakemistoa todennäköisyyksiin perustuvassa tiedonhakujärjestelmässä, kun hakukielenä on voimakkaasti taipuva suomen kieli. Perusmuotoisia hakemistoja oli kaksi, joista toisessa yhdyssanat oli ositettu ja toisessa ei. Tämä siksi, että taivutusmuotohakemistossa yhdyssanoja ei oltu ositettu, ja näin tuloksista saatiin vertailukelpoisia. Tutkimustietokannassa oli käytössä neliportainen relevanssiarviointi ja kyselyjä tehtiin kahdenlaisia: peruskyselyitä ja johdoksilla laajennettuja johdoskyselyitä. Kaikkiin hakemistoihin tehtiin 30 peruskyselyä ja johdoskyselyä. Luonteeltaan tämä tutkimus oli empiirinen evaluointitutkimus. Matemaattista tarkastelua selittämään hakujen erilaista toteutumista hakemistoissa ei tehty.

Tutkimuksessa selvisi, että ositettu perusmuotohakemisto on tiedonhaussa tuloksellisempi kuin osittamaton perusmuotohakemisto tai taivutusmuotohakemisto. Yhdyssanat kannattaa siis tietokantojen käänteistiedostoissa osittaa. Tulos oli samansuuntainen kaikilla tutkituilla relevanssitasoilla. Päätutkimusongelmaan liittyen tarkoituksena oli tutkia, mitä ovat eri hakemistojen heikkoudet ja vahvuudet. Ositetun perusmuotohakemiston vahvuudet ovat selvät: hyvä tuloksellisuus ja helppo käytettävyys. Tiedonhakijan ei tarvitse tuntea sanojen taivutusta, vaan hän voi syöttää hakusanaksi sanan perusmuodon. Osittamattomalla perusmuotohakemistolla ei liene todellista käyttöä, sillä sen tuloksellisuus on heikompi kuin ositetun ja toisaalta kielenanalyysiohjelmissa tuskin saavutetaan säästöä, kun jätetään vain yhdyssanojen ositus pois. Taivutusmuotohakemiston tulokset olivat samaa tasoa kuin osittamattoman perusmuotohakemiston. Sen ainoa vahvuus lienee sen taloudellisuus: taivutusmuotohakemiston muodostaminen on halvempaa, koska ei tarvita perusmuotoistamisohjelmaa eikä yhdyssanojen ositusohjelmaa. Käytännössä sen käyttö Inqueryn kaltaisissa järjestelmissä on todella hankalaa, koska hakuavainten katkaisu ei ole mahdollista. Hakusanan kaikkien taivutusmuotojen lisääminen kyselyyn onnistuu vain kokeellisissa tiedonhakutilanteissa.

Tutkimuksen osaongelmana oli, tuottaako peruskysely vai loogisesti muodostettu johdoskysely parempia tuloksia hakemistojen sisällä. Tulosten mukaan näillä kyselyillä ei ollut tilastollisesti merkitsevää eroa. Myöskään selvää tendenssiä jommankumman kyselyn paremmuudesta ei esiintynyt. Toisena osaongelmana tutkittiin, voidaanko proba-

bilistisessa järjestelmässä taivutusmuotoisessa hakemistossa katkaisua simuloida seulontamenetelmällä eli suorittamalla haku kaikilla hakemistossa esiintyvillä taivutusmuodoilla. Tässä työssä katkaisun simulointi ei ollut täydellistä, koska aidossa katkaisussa mukaan hakuun tulevat kaikki samalla merkkijonolla alkavat sananmuodot, ja tässä sillä vain etsittiin sanan taivutusmuodot. Mekaanisena toimenpiteenä katkaisun simulointi onnistui hyvin ja tulokset olivat lähes samat kuin osittamattomassa perusmuotohakemistossa.

Niiden kielten osalta, joilla on monimutkainen morfologia, on hakemistojen vaikutusta hakujen tuloksellisuuteen on tutkittu aiemmin vain Alkulan (2000) tutkimuksessa. Siinä tiedonhakujärjestelmä perustui täydelliseen täsmäytykseen. Osittaistäsmäytykseen perustuvilla järjestelmillä asiaa ei ole tutkittu. Muissa kielissä tutkimus on suuntautunut lähinnä karsinta-algoritmien (stemmaus) käyttöön kyselyissä eikä niinkään hakemistojen käsittelyyn.

Tämän ja Alkulan (2000) tutkimuksen tutkimusympäristöt vastasivat osittain toisiaan. Tämän tutkimuksen ositettu taivutusmuotohakemisto vastasi Alkulan tutkimusympäristöä 5, jossa haettiin perusmuotoja ja yhdyssanojen kaikkia osia. Edelleen tämän tutkimuksen taivutusmuotohakemisto vastasi Alkulan tutkimusympäristöä 3, joka toteutettiin seulonnalla. Tämän tutkimuksen osittamattomalla perusmuotohakemistolla ei ollut vastinetta Alkulan tutkimuksessa.

Alkulan tutkimuksessa tutkimusympäristöjen 3 ja 5 tarkkuuden paremmuus kyselyissä vaihteli riippuen, mitä operaattoria (JA tai VIRKE) käytettiin, ja tarkasteltiinko tuloksia perusjoukossa, johdososajoukossa vai yhdyssanaosajoukossa. Yleensä näiden kahden tutkimusympäristön väliset erot tarkkuudessa olivat vain pari prosenttiyksikköä, eivätkä ne olleet tilastollisesti merkitseviä. Tässä tutkimuksessa koko kyselyjoukossa ositetun perusmuotohakemiston ja taivutusmuotohakemiston tarkkuuskeskiarvojen erot olivat hieman suurempia, ja tilastollisesti merkitseviä erojakin oli. Kovin suuria nämä erot näiden kahden tutkimusten tuloksien välillä eivät kuitenkaan ole.

Tutkimuksien hakemistot muodostettiin eri tavoin. Alkulan tutkimuksessa hakemistoja luotaessa käytettiin 77:n sanan sulkusanalistaa. Siis myöskään taivutusmuotohakemistoon ei mennyt dokumenttien kaikkia sanoja. Tässä tutkimuksessa käytetyn TUTK-

tietokannan taivutusmuotohakemistoon ovat menneet kaikki dokumenttien sanat. Relevanssiarviointi oli Alkulalla kolmiportainen ja tässä tutkimuksessa neliportainen. Suurin ero tutkimuksissa oli erilaisen tiedonhakujärjestelmän – täystäsmäyttävän ja osittaistäsmäyttävän – käyttö.

Tässä tutkimuksessa hakemistojen sisällä ei perus- ja johdoskyselyn tuloksellisuuden välillä ollut juuri mitään eroa. Alkulan tutkimuksessa erot olivat myös melko pienet, mutta peruskyselyiden tarkkuus oli hieman parempi. Kummassakaan tutkimuksessa ei näillä eroilla ollut tilastollista merkitsevyyttä.

Eri hakemistojen kyselyjen mahdollisimman samanlainen täsmäytyminen dokumentteihin ei aivan onnistunut, koska eri hakemistoissa yhdysmerkin käsittely oli hiukan erilaista. Tämä vaikutti vähän tuloksiin, mutta ei niin paljon, että se estäisi päätelmien teon.

Tutkimus tuotti uutta tietoa informaatiotutkimuksen kentälle, koska siinä tutkittiin agglutinoivan ja synteettisen kielen kannalta tietokannan hakemiston rakennetta, asiaa jota ei ole aiemmin tutkittu. Ei ole syytä epäillä, etteivätkö tulokset olisi yleistettävissä myös muihin agglutinoiviin tai synteettisiin kieliin. Tulokset pätevät etenkin, jos kielessä on runsaasti yhdyssanoja, kuten suomessa.

8 Johtopäätökset

Tämän tutkimuksen tarkoituksena oli selvittää, miten tiedonhaun tuloksellisuus eroaa käytettäessä ositettua perusmuotohakemistoa, osittamatonta perusmuotohakemistoa ja taivutusmuotoista hakemistoa todennäköisyyksiin perustuvassa tiedonhakujärjestelmässä. Erityinen tutkimusongelma oli, kannattaako yhdyssanat osittaa hakemistossa. Tämän selvittämiseksi 30 hakukysymyksestä muodostetut perus- ja johdoskyselyt suoritettiin kaikkiin hakemistoihin. Tulokset laskettiin kolmella relevanssitasolla. Tutkimus rajoittui empiiriseen kokeeseen eikä matemaattista tarkastelua suoritettu.

Tiedonhaun tuloksellisuutta mitattiin laskemalla tarkkuuskeskiarvot vakioiduilla saantitasoilla sekä saanti ja tarkkuus dokumentin katkaisupisteittäin (DCV). Tulosten mukaan ositettu perusmuotoistettu hakemisto on tuloksellisin hakemisto. Hieman huonommin menestyivät osittamaton perusmuotohakemisto ja taivutusmuotohakemisto. Näiden kahden välillä ei ollut eroja. Tulokset olivat samansuuntaiset kaikilla kolmella tutkitulla relevanssitasoilla. Myös täydellisesti täsmäyttävässä ympäristössä perusmuotohakemisto on todettu taivutusmuotoista tuloksellisemmaksi (Alkula, 2000).

Kokotekstitietokantojen, joiden sisällön kielenä on morfologialtaan rikas ja runsaasti yhdyssanoja sisältävä kieli, hakemisto kannattaa siis perusmuotoistaa ja yhdyssanat osittaa, koska 1) hakujen tuloksellisuus on parempi olipa hakujärjestelmä täystäsmäyttävä tai probabilistinen 2) tiedonhakijan on helpompi hakea sanojen perusmuodoilla kuin osata katkaista sanat oikeasta kohdasta, tai kuten osittaistäsmäyttävissä järjestelmissä, syöttää hakulauseeseen kaikki sanan taivutusmuodot.

Lähdeluettelo

Alkula, R. 2000. Merkkijonoista suomen kielen sanoiksi: suomen kielen morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja liittämisen vaikutukset tekstin tallennukseen ja hakuun. Tampereen yliopisto. Acta Universitatis Tamperensis 763. Väitöskirja. Saatavilla www-muodossa:
<URL: http://acta.uta.fi/pdf/951-44-4886-3.pdf>. [Viitattu 8.9.2003].

Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swan, R. and Xu, J. (1998). INQUERY does battle with TREC-6. Julkaisussa Proceedings of the Sixth Text Retrieval Conference (TREC-6). Gaithersburg, MD: National Institute of Standards and Technology, special publication 500-240, 169–206. Saatavilla www-muodossa:
<URL: http://www-2.cs.cmu.edu/~callan/Papers/ir-118.ps.gz>. [Viitattu 8.9.2003].

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

Bookstein, A. 1985. Probability and fuzzy-set applications to information retrieval. Julkaisussa: Williams, M.E. (toim.) Annual Review of Information Science and Technology 20, 117–151.

Broglio, J., Callan, J. P., Croft, W. B. & Nachbar, D. W. 1995. Document retrieval and routing using the INQUERY system. Julkaisussa Proceedings of the Third Text Retrieval Conference (TREC-3). Gaithersburg, MD: National Institute of Standards and Technology, special publication 500-225, 29-38. Saatavilla www-muodossa:
<URL: http://www-2.cs.cmu.edu/~callan/Papers/brogliocallantrec94.ps.gz>. [Viitattu 8.9.2003].

Callan, J. P., Croft, W. B. & Harding, S. M. 1992. The INQUERY Retrieval System. Julkaisussa Proceedings of the Third International Conference on Databases and Expert Systems Applications. Valencia, Spain, 78–83. Saatavilla www-muodossa:
<URL:http://www-2.cs.cmu.edu/~callan/Papers/callancroftdexa92.ps.gz>. [Viitattu 8.9.2003].

Cleverdon, C. W. 1972. On the inverse relationship of recall and precision. Journal of Documentation 28 (3), 195–201.

Conover, W. J. 1980. Practical Nonparametric Statistics. 2nd edition. New York: John Wiley & Sons.

Cooper, W. S. 1994. The formalism of probability theory in IR: a foundation or an encumbrance. Julkaisussa Proceedings of the Seventeenth Annual ACM SIGIR Conference on Research and development in information retrieval. Dublin, Ireland, 242–247.

Cosijn, E. & Ingwersen, P. 2000. Dimensions of relevance. Information Processing and Management 36 (4), 533–550.

Croft, W. B. & Harper, D. J. 1979. Using probabilistic models of document retrieval without relevance information. Journal of Documentation 35 (4), 285–295.

Fidel, R. 1987. Database design for information retrieval: a conceptual approach. New York: Wiley.

Haas, S. W. 1996. Natural language processing: toward large-scale, robust systems. Annual Review of Information Science and Technology 31, 83–119.

Hull, D. 1993. Using statistical testing in the evaluation of retrieval experiments. Julkaisussa Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. Pittsburgh, PA, 329-338. Saatavilla www-muodossa:
<URL: http://www.xrce.xerox.com/people/hull/hull/papers/sigir93.ps>. [Viitattu 8.9.2003].

Iisa, K., Oittinen, H. & Piehl, A. 1999. Kielenhuollon käsikirja. Helsinki: Yrityskirjat.

Ingwersen, P. 1992. Information Retrieval Interaction. London: Taylor Graham.

Inquery document retrieval system. 1996. Ohjetiedosto. Saatavilla www-muodossa:

<URL: http://ciir.cs.umass.edu/irdemo/inqinfo/inqueryhelp.html>. [Viitattu 8.9.2003].

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

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

Keen, E. M. 1992. Presenting results of experimental retrieval comparisons. Information Processing & Management 28 (4), 491–502.

Kekäläinen, J. 1999. The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. University of Tampere. Acta Universitatis Tamperensis 678. Academic dissertation. Saatavilla www-muodossa: <URL: http://www.info.uta.fi/tutkimus/fire/archive/QCES.pdf>. [Viitattu 8.9.2003].

Keskustalo, H. 1994. Suomenkielisen tekstitietokannan hakemiston rakentamisesta INQUERY/TWOL-ympäristössä. Tampereen yliopisto. Informaatiotutkimuksen laitos. Sivuaineen tutkielma.

Kieli ja sen kieliopit: opetuksen suuntaviivoja. 2000. Helsinki: Painatuskeskus.

Koskenniemi, K. 1985. Finstems: a module for information retrieval. Julkaisussa: Karlsson, F. (ed.) Computational morphosyntax: report on research 1981–84. University of Helsinki. Publications of the Department of General Linguistics 13. S. 81–92.

Leino, P. 1991. Kieleen mieltä: hyvää suomea. Helsinki: Otava.

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

Mellin, I. 1996. Johdatus tilastotieteeseen. 2. kirja. Tilastotieteen jatkokurssi. Helsingin yliopisto, tilastotieteen laitos.

Mizzaro, S. 1997. Relevance: the whole history. Journal of the American Society for Information Science 48 (9), 810–832.

Pirkola, A. 1999. Studies on linguistic problems and methods in text retrieval. University of Tampere. Acta Universitatis Tamperensis 672. Academic dissertation.

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

Robertson, S.E. 1977. The probability ranking principle in IR. Journal of Documentation 33 (4), 294–304.

Salton, G. & McGill M. J. 1984. Introduction to Modern Information Retrieval. International student edition. Auckland: McGraw-Hill International Book Company.

Salton, G. 1992. The state of retrieval system evaluation. Information Processing & Management 28 (4), 441–449.

Saracevic, T. 1975. Relevance: a review of and a framework for thinking on the notion in information science. Journal of the American Society for Information Science 26 (6), 321–343.

Saracevic, T. 1996. Relevance reconsidered '96. Julkaisussa Ingwersen, P. and Pors, P. O. eds. Proceedings of the Second International Conference on Conceptions of Library and Information Science: Integration in Perspective. Copenhagen: The Royal School of Librarianship, 201–218.

Siegel, S. & Castellan, N. J. 1988. Nonparametric Statistics for the Behavioral Sciences. 2nd edition. New York: McGraw-Hill Book Company.

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

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

Su, L.T. 1992. Evaluation measures for interactive information retrieval. Information Processing & Management 28 (4), 503–516.

Suomen sanojen alkuperä : etymologinen sanakirja, osa 2. 1995. Helsinki: Suomalaisen Kirjallisuuden Seura.

Tietotekniikan sanasto. 1990. Helsinki: Tietosanoma.

Turtle, HR. 1991. Inference networks for document retrieval. University of Massachusetts, Department of Computer and Information Science. Saatavilla www-muodossa: <URL:http://citeseer.nj.nec.com/turtle91inference.html>. [Viitattu 8.9.2003].

Van Rijsbergen, C.J. 1979. Information retrieval. 2nd ed. London: Butterworths.

Vesikansa, J. 1977. Johdokset. (Nykysuomen oppaita 2). Porvoo: WSOY.

LIITE 1: Hakukysymykset

Hakuaiheet, käsitetyypit ja saantikantojen koko eri relevanssitasoilla.

Aihe		Käsitetyyppi	Liberaali rel.taso	Normaali rel.taso	Tiukka rel.taso
1	George Bushin ja Mihail Gorbatshovin tapaaminen Helsingissä syyskuussa 1990. Neuvotteluissa käsitellyt asiat sekä tehdyt päätökset ja sopimukset.	henkilö	54	32	14
2	Etelä-Amerikan velkakriisi. Miten velkaantumisongelma on kehittynyt? Miten ongelmaa on pyritty ratkaisemaan?	maantiet. rajattu aihe	81	55	11
3	Metsäteollisuuden polkumyyntisyytteet USA:ssa. Kiinnostavaa suomalaisten paperinviejien kohtalo. Polkumyyntisyytösten sisältö, oikeudenkäyntien tulokset.	aihe	21	16	10
4	Jyväskylän kaupungin ja maalaiskunnan kuntaliitoshanke. Halutaan kartoittaa liitoshankkeen kannattajien ja vastustajien mielipiteitä ja perusteluja. Arviot liitoksen taloudellisista vaikutuksista (mm. porkkanaraha).	maantiet. rajattu aihe	16	8	7
6	Varsovan liiton lakkauttaminen. Mitä tahansa muutosprosessista, eri jäsenmaiden suhtautumisesta, päätöksistä, jne.	organisaatio	129	48	17
7	Neuvostoliiton Liettuaan kohdistama taloussaarto keväällä 1990. Mitä toimia taloussaartoon liittyi ja miten se näkyi Liettuassa? Saarron lopettamiseen johtaneet tapahtumat.	maantiet. rajattu aihe	145	87	15
8	Irakin joukkotuhoaseiden hävittäminen. Irakin on Persianlahden sodan aselepsopimuksen mukaan luovuttava kemiallisista, biologisista ja ydinaseista ja niiden tuotantotekniikasta. YK vastaa aseiden inventoinnista ja hävittämisestä. Miten tehtävän suoritus on onnistunut.	maantiet. rajattu aihe	83	65	39
9	OPEC:n öljyn hintaa ja tuotantomääriä koskevat päätökset.	organisaatio	72	29	6
10	Presidentti Iliescun hallituksen avuksi kutsumien kaivosmiesten väkivaltaisuudet oppositiota vastaan Bukarestissa. Taustatietoja tapahtumista, uhreista ja jälkiselvittelyistä.	maantiet. rajattu aihe	37	24	7
11	Namibian itsenäistymiseen liittynyt YK:n rauhanturvaoperaatio. Tietoja operaation valmistelusta, siihen liittyneistä tapahtumista sekä UNTAG-joukkojen ja sen suomalaispataljoonan toiminnasta.	maantiet. rajattu aihe	143	98	47
12	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.	organisaatio	54	29	13
13	Carl Bildt ja pohjoismainen yhteistyö. Bildtin pohjoismaista yhteistyötä koskevat	henkilö	19	13	1

	lausunnot. Mitä erityistä Bildt on sanonut Ruotsin ja Suomen yhteistyöstä?				
14	Jugoslavian presidenttineuvoston toimintaa koskevat uutiset. Erityisesti tiedot istunnoista ja niissä tehdyistä päätöksistä.	organisaatio	111	36	19
15	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?	maantiet. rajattu aihe	84	54	17
17	Valmetin traktori- ja kuljetusvälinetuotannon kannattavuus. Kuljetusvälinetoimialaan lasketaan kuuluviksi metsä- ja siirtokoneet sekä kiskokalusto (mm. Transtech). Osakkuudet henkilö- ja kuorma-autoteollisuudessa jätetään tarkastelun ulkopuolelle.	organisaatio	68	45	6
20	Tampellan irtisanomiset. Tavoitteena koota tietoja Tampella-konserniin kuuluvien yhtiöiden suorittamista irtisanomisista. Tietoja lomautuksista ja lyhennetyistä työviikoista ei tarvita.	organisaatio	26	14	1
21	Keran ja KTM:n investoinnit matkailuun. Tietoja matkailualan yrityksille myönnetyistä avustuksista ja lainoista (= tässä investointi). Erityisen arvokkaita yhteenvedot.	organisaatio	33	17	2
22	Neste oy:n maakaasutoiminta. Halutaan yleiskuva Nesteen maakaasutoiminnoista. Mitä Neste on puuhaillut maakaasun hankinnan (kentät ja tuontisopimukset), jakelun (verkoston rakentaminen) ja markkinoinnin alueilla.	organisaatio	65	37	10
23	Ydinvoimalaitosten tuottamien radioaktiivisten jätteiden käsittely ja varastointi. Esimerkkejä ongelmista, riskeistä ja sattuneista ydinjätevahingoista.	aihe	68	34	26
24	AIDSin levinneisyys EY-maissa. Miten vakava AIDS-tilanne on näissä maissa? Tietoja esiintymämääristä ja kampanjoista ym. taudin leviämistä ehkäisevistä toimista.	maantiet. rajattu aihe	43	21	8
25	Elintarvikkeiden tuontirajoitukset ja -säännöstely eri maissa. Rajasuojan ja sen vähentämisen vaikutus elintarviketeollisuuteen erityisesti Suomessa. Selvityksiä, arvioita, mielipiteitä ym. taustatietoa.	aihe	90	14	3
26	Asuntotuotannon suhdanteet ja suhdannevaihtelut Suomessa (valtakuntataso); erityisesti tilasto- ja ennustetietoja, arvioita (rakentamisesta ei asuntokaupasta).	aihe	122	36	3
27	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.	aihe	136	87	17
28	Japanin autoteollisuuden investoinnit Eu-	maantiet.	24	16	5

	rooppaan ja tuotannollinen yhteistyö eurooppalaisten autonvalmistajien kanssa. Mihin maihin japanilaisia autotehtaita on suunniteltu, perustettu ja laajennettu? Tuotantomäärät ja -trendit.	rajattu aihe			
29	Metsäteollisuuden ympäristöinvestoinnit. Rajoitutaan vesiensuojeluun liittyviin investointeihin kemiallisessa metsäteollisuudessa. Sekä varsinaiset puhdistamoinvestoinnit että ympäristöystävällisempien prosessien käyttöönotto.	aihe	45	25	13
30	Kaupan aukioloajat. Halutaan selvittää vähittäiskauppojen aukioloaikojen vapauttamista koskevaa keskustelua. Erityisesti kartoitetaan kaupan järjestöjen ja ammattijärjestöjen kannanottoja ja toimia.	aihe	35	27	13
31	Pakkaukset ympäristönsuojelukysymyksenä. Erityisesti kiinnostavat kulutustavarapakkausten kierrätysjärjestelmät, niiden kehittämiskokeilut, kierrätykseen liittyvä lainsäädäntö eri maissa.	aihe	73	59	26
33	Esko Ahon ja Suomen EY-jäsenhakemus. Ahon Suomen EY-jäsenyyden hakemiseen liittyvät mielipiteet, kannanotot ja toimet. Muiden arviot Eskon toimista ja puheista.	henkilö	35	21	6
34	Kauko Juhantalon ydinvoimapuheet ja -teot. Juhantalon perustelut 5. ydinvoimalan puolesta. Miten Juhantalo vei ydinvoimaratkaisua eteenpäin?	henkilö	14	6	2
35	Vihreiden tekemät aloitteet, välikysymykset, ehdotukset, puheenvuorot ja äänestyskäyttäytyminen Suomen eduskunnassa. Tarkastelussa sekä ryhmä että yksittäiset kansanedustajat.	organisaatio	27	13	2
		yhteensä	1953	1066	366
		keskiarvo	65,1	35,5	12,2

LIITE 2: Kyselyt

Peruskyselyt ositettuun perusmuotohakemistoon

#q1 = #sum(george bush mihail gorbatshov tapaaminen helsinki syyskuu 1990 neuvottelu asia p{{t|s sopimus);

#q2 = #sum(etel{-amerikka velkakriisi velkaantumisongelma kehitty{ ongelma pyrki{ ratkaista);

#q3 = #sum(mets{teollisuus polkumyyntisyyte usa kiinnostaa suomalainen paperinviej{ kohtalo polkumyyntisyyt|s sis{lt| oikeudenk{ynti tulos);

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

#q6 = #sum(varsoa varsova liitto lakkauttaminen muutosprosessinen muutosprosessi eri j{senmaa suhtautuminen p{{t|s);

#q7 = #sum(neuvostoliitto liettua kohdistaa taloussaarto kev{t 1990 toimi toimia liitty{ n{ky{ saarto lopettaminen johtaa tapahtua tapahtuma);

#q8 = #sum(irak joukkotuhoase h{vitt{minen persianlahti sota aseleposopimus luovuttaa luopua kemiallinen biologinen ydinase ydinaseinen tuotantotekniikka yk vastata ase inventointi teht{v{ suoritus onnistua);

#q9 = #sum(opec |ljy hinta tuotantom{{ri tuotantom{{r{ koskea p{{t|s);

#q10 = #sum(presidentti @iliescun hallitus apu avu kaivosmies v{kivaltaisuus oppositio vastata bukarest taustatieto tapahtua tapahtuminen tapahtuma uhri j{lkiselvittely);

#q11 = #sum(namibia itsen{istyminen liitty{ yk rauhanturvaoperaatio tieto operaatio valmistelu tapahtua tapahtuminen tapahtuma #0(@untag /joukko) suomalaispataljoona toiminta);

#q12 = #sum(ey parlamentti asema yhteis| p{{t|ksenteko selvitt{{ suhde komissio toimielin muutos nykyinen demokraattinen kontrolli toimia);

#q13 = #sum(carl @bildt pohjoismainen yhteisty| @bildtin pohjoismaa koskea lausunto erit{ sanoa ruotsi suomi);

#q14 = #sum(jugoslavia presidenttineuvosto toiminta koskea uutinen tieto istunto istunta p{{t|s);

#q15 = #sum(l{nsi-saksa it{-saksa miehitt{j{valtio yhdysvallat yhdysvalta iso -britannia ranska neuvostoliitto #1(2 4 /neuvottelu) saksa yhdistyminen ratkaista kysymys riitakysymys olennainen synty{ sopimus);

#q17 = #sum(valmet traktorituotanto kuljetusv{linetuotanto kannattavuus kuljetusv{linetoimiala kuulupa mets{kone siirtokone kiskokalusto @transtech osakkuus henkil|autoteollisuus kuorma-autoteollisuus j{tt{{ tarkastelu);

#q20 = #sum(tampella irtisanominen tavoite koota tieto tampella-konserni yhti| suorittaa suorittaminen lomautus lyhent{{ ty|viikko);

#q21 = #sum(kera #1(@ktm n) investointi matkailu tieto matkailuala yritys my|nt{{ avustus laina arvokas yhteenveto);

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

#q23 = #sum(ydinvoimalaitos tuottaa radioaktiivinen j{te k{sittely varastointi esimerkki ongelma riski sattua ydinj{tevahinko);

#q24 = #sum(aids levinneisyys ey-maa vakava aids-tilanne aids-tila maa tieto esiintym{m{{r{ esiintym{m{{ri kampanja tauti levi{minen levit{ ehk{ist{ toimi);

#q25 = #sum(elintarvike tuontirajoitus tuontis{{nn|stely eri maa rajasuoja v{hent{minen vaikuttua vaikutus elintarviketeollisuus suomi selvitys arvioittaa arvio mielipide taustatieto);

#q26 = #sum(asuntotuotanto suhdanne suhdannevaihtelu suomi valtakuntataso tilastotieto ennustetieto arvio arvioittaa rakentaminen asuntokauppa);

#q27 = #sum(tieliikenne p{{st| suomi ulkomaa kehitty{ odottaa lains{{d{nt| vaikuttua vaikutus merkitt{{ merkit{ katalysaattori yleistyminen vaikuttaa p{{st|taso katalysaattoritekniikka sin{ns{ kiinnostaa);

#q28 = #sum(japani autoteollisuus investointi eurooppa tuotannollinen yhteisty| eurooppalainen autonvalmistaja maa japanilainen autotehdas suunnitella perustaa laajentaa tuotantom{{r{ tuotantotrendi);

#q29 = #sum(mets{teollisuus ymp{rist|investointi rajoittua vesiensuojelu liitty{ investointi kemiallinen varsinainen puhdistamoinvestointi ymp{rist|yst{v{llinen prosessi k{ytt||notto);

#q30 = #sum(kauppa aukioloaika selvitt{{ v{hitt{iskauppa vapauttaa vapauttaminen koskea keskustelu kartoittaa j{rjest| ammattij{rjest| kannanotto toimia toimi);

#q31 = #sum(pakkaus ymp{rist|nsuojelukysymys kiinnostaa kulutustavarapakkaus kierr{tysj{rjestelm{ kehitt{miskokeilu kierr{tys liitty{ lains{{d{nt| eri maa);

#q33 = #sum(esko aho suomi ey-j{senhakemus ey-j{senyys hakeminen liitty{ mielipide kannanotto toimi arvio puhe puh);

#q34 = #sum(kauko juhantalo ydinvoimapuhe ydinvoimateko perustelu 5 ydinvoimala vied{ ydinvoimalaratkaisu);

#q35 = #sum(vihre{ aloite v{likysymys ehdotus puheenvuoro {{nestysk{ytt{ytyminen suomi eduskunta tarkastelu ryhm{ yksitt{inen kansanedustaja);

Johdoskyselyt ositettuun perusmuotohakemistoon

#q1 = #sum (george bush mihail gorbatshov #syn(tapaaminen tavata tapaaja) #syn(helsinki helsinkil{inen) #syn(syyskuu syyskuinen) 1990 #syn(neuvottelu neuvotella neuvottelija neuvotteleminen) #syn(asia asiallinen asiaton) p{{t|s #syn(sopimus sopimuksellinen sopimuksetон));

#q2 = #sum(#syn(etel{-amerikka etel{amerikkalainen) velkakriisi velkaantumisongelma #syn(kehitty{ kehittyminen kehitt{{ kehitt{j{ kehitt{minen kehitell{) #syn(ongelma ongelmainen ongelmallinen ongelmaton) #syn(pyrki{ pyrkij{ pyrkiminen) #syn(ratkaista ratkaisija ratkaiseminen));

#q3 = #sum(mets{teollisuus polkumyyntisyyte usa #syn(kiinnostaa kiinnostua) suomalainen paperinviej{ #syn(kohtalo kohtaloinen kohtalollinen) polkumyyntisyyt|s #syn(sis{lt| sis{lt|inen sis{ll|llinen) oikeudenk{ynti #syn(tulos tuloksinen tuloksellinen tuloksetон));

#q4 = #sum(#syn(jyv{skyl{ jyv{skyl{l{inen) #syn(kaupunki kaupunkinen kaupunkilainen kaupungillinen) #syn(maalaiskunta maalaiskuntalainen) kuntaliitoshanke #syn(kartoittaa kartoittaja kartoittaminen) liitoshanke #syn(kannattaja kannattaa kannattaminen kannatella) #syn(vastustaja vastustaa vastustaminen vastustella) mielipide perusteluja #syn(perustelu perusteluinen perustella perusteleminen) arvio #syn(liitos liitoksinen) #syn(taloudellinen talous) #syn(vaikutus vaikutuksinen vaikutukseton) porkkanaraha);

#q6 = #sum(#syn(varsoa varsoja) #syn(varsova varsovalainen) #syn(liitto liittolainen) #syn(lakkauttaminen lakkauttaa lakkautella lakkautua) muutosprosessinen muutosprosessi #syn(eri erilainen erillinen eritt{in) j{senmaa #syn(suhtautuminen suhtautua) p{{t|s);

#q7 = #sum(#syn(neuvostoliitto neuvostoliittolainen) #syn(liettua liettualainen) #syn(kohdistaa kohdistaminen kohdistua) talous-saarto #syn(kev{t kev{inen kev{{llinen) 1990 #syn(toimi toiminen toimellinen toimetoн) #syn(toimia toimija toimiminen) #syn(liitty{ liittyj{ liittyminen) #syn(n{ky{ n{kyj{ n{kyminen) saarto #syn(lopettaminen lopettaa lopettaja lopetella) #syn(johtaa johtaja johtaminen johtua) #syn(tapahtua tapahtuminen) tapahtuma);

#q8 = #sum(#syn(irak irakilainen) #syn(joukkotuhoase joukkotuhoaseinen) #syn(h{vitt{minen h{vitt{{ h{vitt{j{) persianlahti #syn(sota sodaton) aseleposopimus #syn(luovuttaa luovuttaja luovuttaminen) #syn(luopua luopuja luopuminen) #syn(kemiallinen kemia) biologinen #syn(ydinase ydinaseeton) ydinaseinen tuotantotekniikka yk #syn(vastata vastaaja vastaaminen vastailla) #syn(ase aseinen aseellinen aseeton) #syn(inventointi inventoida inventoija inventoiminen) teht{v{ #syn(suoritus suorittaa suorittaja suorittaminen) #syn(onnistua onnistuja onnistuminen onnistaa));

#q9 = #sum(opec #syn(|ljy |ljyinen |ljyt|n |ljyt{) #syn(hinta hintainen hinnallinen) tuotantom{{ri tuotantom{{r{ #syn(koskea koskeminen) p{{t|s);

#q10 = #sum(#syn(presidentti presidentillinen presidentit|n) @iliescun hallitus #syn(apu apulainen avullinen avuton) avu #syn(kaivosmies kaivosmiehinen) #syn(v{kivaltaisuus v{kivalta v{kivaltainen v{kivallaton) oppositio #syn(vastata vastaaja vastaaminen vastailla) #syn(bukarest bukarestilainen) taustatieto tapahtua tapahtuminen tapahtuma uhri j{lkiselvittely);

#q11 = #sum(#syn(namibia namibialainen) #syn(itsen{istyminen itsen{isty{) #syn(liitty{ liittyj{ liittyminen) yk rauhanturvaoperaatio #syn(tieto tietoinen tiedollinen tiedoton) operaatio #syn(valmistelu valmisteluinen valmistella valmistelija valmisteleminen) tapahtua tapahtuminen tapahtuma #0(@untag /joukko) suomalaispataljoona #syn(toiminta toiminnallinen toiminnoittain toimia toimija toimiminen));

#q12 = #sum(ey #syn(parlamentti parlamenttinen) #syn(asema asemallinen) #syn(yhteis| yhteis|inen yhteis|llinen) p{{t|ksenteko #syn(selvitt{{ selvitt{j{ selvitt{minen selvit{ selvi{minen selvitell{ selviyty{) #syn(suhde suhteinen suhteellinen suhteeton) komissio #syn(toimielin toimieliminen) muutos nykyinen #syn(demokraattinen demokraatti) #syn(kontrolli kontrollinen) #syn(toimia toimija toimiminen));

#q13 = #sum(carl @bildt pohjoismainen yhteisty| @bildtin #syn(pohjoismaa pohjoismaalainen) #syn(koskea koskeminen) lausunto erit{ #syn(sanoa sanoja sanominen sanella sanoutua) #syn(ruotsi ruotsalainen) #syn(suomi suominen suomalainen));

#q14 = #sum(#syn(jugoslavia jugoslavialainen) presidenttineuvosto #syn(toiminta toiminnallinen toiminnoittain toimia toimija toimiminen) #syn(koskea koskeminen) uutinen #syn(tieto tietoinen tiedollin en tiedoton) istunto #syn(istunta istua istuja istuminen istuskella istuutua) p{{t|s);

#q15 = #sum(#syn(l{nsi-saksa l{nsisaksalainen) #syn(it{-saksa it{saksalainen) miehitt{j{valtio #syn(yhdysvallat yhdysvaltalainen) yhdysvalta iso-britannia #syn(ranska ranskalainen) #syn(neuvostoliitto neuvostoliittolainen) #1(2 4 #syn(/neuvottelu /neuvotella /neuvottelija /neuvotteleminen)) #syn(saksa saksalainen) #syn(yhdistyminen yhdisty{) #syn(ratkaista ratkaisija ratkaiseminen) kysymys riitakysymys olennainen #syn(synty{ syntyj{ syntyminen) #syn(sopimus sopimuksellinen sopimukseton));

#q17 = #sum(valmet traktorituotanto kuljetusv{linetuotanto #syn(kannattavuus kannattava) kuljetusv{linetoimiala kuulupa #syn(mets{kone mets{koneinen) siirtokone kiskokalusto @transtech #syn(osakkuus osakas) henkil|autoteollisuus kuorma-autoteollisuus #syn(j{tt{{ j{tt{j{ j{tt{minen j{tt{yty{) #syn(tarkastelu tarkastella tarkastelija));

#q20 = #sum(#syn(tampella tampellalainen) #syn(irtisanominen irtisanoa) #syn(tavoite tavoitteinen tavoitteellinen tavoitteeton) #syn(koota kokoaja kokoaminen kokoilla) #syn(tieto tietoinen tiedollinen tiedoton) tampella-konserni #syn(yhti| yhti|itt{in) #syn(suorittaa suorittaja) suorittaminen #syn(lomautus lomauttaa lomauttaminen) #syn(lyhent{{ lyhent{minen lyhennell{ lyhenty{ lyhet{) ty|viikko);

#q21 = #sum(kera #1(@ktm n) #syn(investointi investoida investoija investoiminen) #syn(matkailu matkailullinen matkailla matkailija) #syn(tieto tietoinen tiedollinen tiedoton) matkailuala #syn(yritys yrityksitt{in yritt{{ yritt{j{ yritt{minen yritell{) #syn(my|nt{{ my|nt{j{ my|nt{minen my|nnell{ my|nty{) #syn(avustus avustaa avustaja avustaminen) #syn(laina lainaton lainata) arvokas yhteenveto);

#q22 = #sum(#syn(neste nesteinen) maakaasutoiminta yleiskuva maakaasutoiminto #syn(puuhailla puuhailija puuhaileminen puuhata puuhaaja puuhaaminen) maakaasu #syn(hankinta hankkia hankkija hankkiminen hankkiutua) #syn(kentt{ kent{llinen) tuontisopimus #syn(jakelu jaella jakelija jakeleminen) verkosto #syn(rakentaminen rakentaa rakentaja rakennella) #syn(markkinointi markkinoinnillinen markkinoida markkinoija markkinoiminen) #syn(alue alueinen alueellinen alueittain));

#q23 = #sum(ydinvoimalaitos #syn(tuottaa tuottaja tuottaminen) #syn(radioaktiivinen radioaktiivi) j{te #syn(k{sittely k{sitell{ k{sittelij{ k{sitteleminen) #syn(varastointi varastoida varastoiminen) #syn(esimerkki esimerkillinen) #syn(ongelma ongelmainen ongelmallinen ongelmaton) #syn(riski riskillinen riskit|n) #syn(sattua sattuminen) ydinj{tevahinko);

#q24 = #sum(aids levinneisyys #syn(ey-maa ey-maittain) vakava aids-tilanne aids-tila #syn(maa mainen maalainen maallinen maaton maittain) #syn(tieto tietoinen tiedollinen tiedoton) esiintym{m{{r{ esiintym{m{{ri kampanja #syn(tauti tautinen) levi{minen levit{ #syn(ehk{ist{ ehk{isij{ ehk{iseminen) #syn(toimi toiminen toimellinen toimeton));

#q25 = #sum(elintarvike tuontirajoitus tuontis{{nn|stely #syn(eri erilainen erillinen eritt{in) #syn(maa mainen maalainen maallinen maaton maittain) rajasuoja #syn(v{hent{minen v{hent{{ v{hent{j{) #syn(vaikuttua vaikuttaa vaikuttaja vaikuttaminen) #syn(vaikutus vaikutuksinen vaikutukseton) elintarviketeollisuus #syn(suomi suominen suomalainen) #syn(selvitys selvitt{{ selvitt{j{ selvitt{minen selvitell{) arvioittaa arvio mielipide taustatieto);

#q26 = #sum(asuntotuotanto suhdanne suhdannevaihtelu #syn(suomi suominen suomalainen) valtakuntataso tilastotieto ennustetieto arvio arvioittaa #syn(rakentaminen rakentaa rakentaja rakennella rakentua) asuntokauppa);

#q27 = #sum(tieliikenne #syn(p{{st| p{{st|t|n) #syn(suomi suominen suomalainen) #syn(ulkomaa ulkomainen ulkomaalainen) #syn(kehitty{ kehittyminen kehitt{{ kehitt{j{ kehitt{minen kehitell{) #syn(odottaa odottaja odottaminen odotella) #syn(lains{{d{nt| lains{{d{nn|llinen) vaikuttua #syn(vaikutus vaikutuksinen vaikutukseton) merkitt{{ #syn(merkit{ merkitsij{ merkitseminen) #syn(katalysaattori katalysaattoriton) #syn(yleistyminen yleisty{) #syn(vaikuttaa vaikuttaja vaikuttaminen) p{{st|taso katalysaattoritekniikka sin{ns{ kiinnostaa);

#q28 = #sum(japani autoteollisuus #syn(investointi investoida investoija investoiminen) eurooppa #syn(tuotannollinen tuotanto) yhteisty| eurooppalainen autonvalmistaja #syn(maa mainen maalainen maallinen maaton maittain) japanilainen autotehdas #syn(suunnitella suunnittelija suunnitteleminen) #syn(perustaa perustaja perustaminen perustella) #syn(laajentaa laajentaja laajentaminen laajentua laajeta) tuotantom{{r{ tuotantotrendi);

#q29 = #sum(mets{teollisuus ymp{rist|investointi #syn(rajoittua rajoittumin en rajoittaa rajoittaja rajoittaminen) vesiensuojelu #syn(liitty{ liittyj{ liittyminen) #syn(investointi investoida investoija investoiminen) #syn(kemiallinen kemia) varsinainen puhdistamoinvestointi ymp{rist|yst{v{llinen #syn(prosessi prosessinen) k{ytt|notto);

#q30 = #sum(#syn(kauppa kaupallinen kaupaton) aukioloaika #syn(selvitt{{ selvitt{j{ selvitt{minen selvitell{ selvit{ selvi{minen selviyty{) v{hitt{iskauppa #syn(vapauttaa vapauttaja vapautella) vapauttaminen #syn(koskea koskeminen) #syn(keskustelu keskustella keskustelija keskusteleminen) #syn(kartoittaa kartoittaja kartoittaminen) #syn(j{rjest| j{rjest|llinen) #syn(ammattij{rjest| ammattij{rjest|llinen) kannanotto #syn(toimia toimija toimiminen) #syn(toimi toiminen toimellinen toimeton));

#q31 = #sum(#syn(pakkaus pakata) ymp{rist|nsuojelukysymys #syn(kiinnostaa kiinnostua) kulutustavarapakkaus kierr{tysj{rjestelm{ kehitt{miskokeilu #syn(kierr{tys kierr{tt{{ kierr{tt{j{ kierr{tt{minen) #syn(liitty{ liittyj{ liittyminen) #syn(lains{{d{nt| lains{{d{nn|llinen) #syn(eri erilainen erillinen eritt{in) #syn(maa mainen maalainen maallinen maaton maittain));

#q33 = #sum(esko #syn(aho aholainen) #syn(suomi suominen suomalainen) ey-j{senhakemus ey-j{senyys #syn(hakeminen hakea hakija hakeutua) #syn(liitty{ liittyj{ liittyminen) mielipide kannanotto #syn(toimi toiminen toimellinen toimeton) arvio #syn(puhe puheinen) puh);

#q34 = #sum(kauko juhantalo ydinvoimapuhe ydinvoimateko #syn(perustelu perusteluinen perustella perusteleminen) 5 ydinvoimala #syn(vied{ viej{ vieminen) ydinvoimalaratkaisu);

#q35 = #sum(#syn(vihre{ vihre{llinen) #syn(aloite aloitteinen aloitteellinen) v{likysymys #syn(ehdotus ehdottaa ehdottaja ehdottaminen) #syn(puheenvuoro puheenvuoroinen) {{nestysk{ytt{ytyminen #syn(suomi suominen suomalainen) eduskunta #syn(tarkastelu tarkastella tarkastelija) #syn(ryhm{ ryhm{l{inen ryhmitt{in) yksitt{inen kansanedustaja);

Perus- ja johdoskyselyt osittamattomaan perusmuotohakemistoon ovat muuten samanlaiset kuin ositettuun perusmuotohakemistoon, paitsi, että jokaisen hakuavaimen edessä on vinoviiva (/).

Peruskyselyt taivutusmuotohakemistoon (esimerkki hakukysymyksestä numero 2)

#q2 = #sum(#syn(etelä-amerikan etelä-amerikassa etelä-amerikassakin etelä-amerikasta etelä-amerikka etelä-amerikkaa etelä-amerikkaan etelä-amerikkakaan) #syn(velkakriisi velkakriisiin velkakriisin velkakriisistä velkakriisiä) #syn(velkaantumisongelma velkaantumisongelman) #syn(kehittyessä kehittyessään kehittyi kehittyisi kehittyisikö kehittyisivät kehittyivät kehittymistä kehittymistäkin kehittymistään kehittymässä kehittymästä kehittymään kehittynee kehittyneeksi kehittyneelle kehittyneellä kehittyneemmillä kehittyneemmiltä kehittyneemmissä kehittyneemmälle kehittyneemmän kehittyneemmät kehittyneempi kehittyneempien

kehittyneempikin kehittyneempiä kehittyneempänä kehittyneempää kehittyneempään kehittyneen kehittyneenä kehittyneeseen kehittyneessä kehittyneestä kehittyneet kehittyneiden kehittyneille kehittyneillä kehittyneiltä kehittyneimmille kehittyneimmistä kehittyneimmän kehittyneimmät kehittyneimpien kehittyneimpiin kehittyneimpiä kehittyneimpään kehittynein kehittyneintä kehittyneisiin kehittyneissä kehittyneissäkään kehittyneistä kehittyneitten kehittyneitä kehittynevät kehittynyt kehittynytkin kehittynyttä kehittyvien kehittyviin kehittyville kehittyvillä kehittyvine kehittyvissä kehittyvistä kehittyviä kehittyvä kehittyväksi kehittyvälle kehittyvällä kehittyvän kehittyvänkin kehittyvänä kehittyvässä kehittyvästä kehittyvät kehittyvää kehittyvään kehittyy kehittyä kehittyäkseen kehity kehitykään kehityttyä kehityttävä) #syn(ongelma ongelmaa ongelmaamme ongelmaan ongelmaansa ongelmahan ongelmakseen ongelmaksi ongelmalle ongelmalta ongelmamme ongelman ongelmana ongelmanaan ongelmanamme ongelmani ongelmanne ongelmansa ongelmassa ongelmasta ongelmastaan ongelmat ongelmia ongelmiaan ongelmiahan ongelmiakaan ongelmiakin ongelmiamme ongelmianne ongelmiemme ongelmien ongelmiensa ongelmiimme ongelmiin ongelmiinsa ongelmiksi ongelmilla ongelmille ongelmilta ongelmina ongelminaan ongelmineen ongelmissa ongelmissaan ongelmissakin ongelmista ongelmistaan ongelmistani ongelmitta) #syn(pyrityn pyrimmekin pyrimme pyrikään pyri pyrittävän pyrittäväksi pyrittävä pyrittäneen pyrittäisiinkin pyrittäisiin pyrin pyritte pyrittiin pyritty pyrittykään pyrittyään pyrittäessä pyrittäisi pyritä pyritäkään pyritään pyritäänhän pyritäänkin pyritäänkö pyrki pyrkien pyrkiessä pyrkiessämme pyrkiessään pyrkii pyrkiikin pyrkiikö pyrkikin pyrkikäämme pyrkikö pyrkimistä pyrkimällä pyrkimässä pyrkimättä pyrkimään pyrkinee pyrkineekin pyrkineelle pyrkineen pyrkineensä pyrkineessä pyrkineet pyrkineetkään pyrkineiden pyrkineiltä pyrkineistä pyrkineitä pyrkinevät pyrkinyt pyrkinytkin pyrkinytkään pyrkinyttä pyrkisi pyrkisikö pyrkisimme pyrkisin pyrkisivät pyrkivien pyrkiviin pyrkiville pyrkivillä pyrkiviltä pyrkivinä pyrkivissä pyrkivistä pyrkiviä pyrkivä pyrkiväksi pyrkivälle pyrkivällä pyrkivältä pyrkivän pyrkivänsä pyrkivänä pyrkivässä pyrkivästä pyrkivät pyrkivätkin pyrkivää pyrkivään pyrkiä pyrkiäkin pyrkiäkseen) #syn(ratkaise ratkaisee ratkaiseekin ratkaisemaan ratkaisemalla ratkaisemassa ratkaisematta ratkaisemiksi ratkaisemin ratkaisemista ratkaisemistaan ratkaisemme ratkaiseva ratkaisevaa ratkaisevaan ratkaisevaksi ratkaisevaksikin ratkaisevalla ratkaisevalle ratkaisevalta ratkaisevammaksi ratkaisevammassa ratkaisevammin ratkaisevampaa ratkaisevampi ratkaisevan ratkaisevana ratkaisevani ratkaisevansa ratkaisevassa ratkaisevasta ratkaisevasti ratkaisevastikin ratkaisevat ratkaisevia ratkaisevien ratkaiseviin ratkaiseviksi ratkaiseville ratkaisevimmaksi ratkaisevimmassa ratkaisevimmat ratkaisevimmin ratkaisevimpaan ratkaisevimpana ratkaisevimpiin ratkaisevin ratkaisevina ratkaisevinta ratkaisevista ratkaisi ratkaisimme ratkaisisi ratkaisisivat ratkaisivat ratkaiskaa ratkaiskoon ratkaisse ratkaissee ratkaisseemme ratkaisseen ratkaisseensa ratkaisseet ratkaisseista ratkaissut ratkaista ratkaistaan ratkaistaanko ratkaistaessa ratkaistaisiin ratkaistakaan ratkaistakseen ratkaistaksemme ratkaistane ratkaistaneen ratkaistava ratkaistavaa ratkaistavakseen ratkaistavaksi ratkaistavalta ratkaistavan ratkaistavana ratkaistavanaan ratkaistavat ratkaistavia ratkaistaviin ratkaistaviksi ratkaistavina ratkaistavissa ratkaistavista ratkaistessa ratkaistessaan ratkaistiin ratkaistu ratkaistua ratkaistuaan ratkaistuiksi ratkaistuja ratkaistuksi ratkaistun ratkaistuna ratkaistusta ratkaistut));

Johdoskyselyt taivutusmuotohakemistoon (esimerkki hakukysymyksestä numero 2)

#q2 = #sum(#syn(etelä-amerikan etelä-amerikassa etelä-amerikassakin etelä-amerikasta etelä-amerikka etelä-amerikkaa eteläamerikkaan etelä-amerikkakaan eteläamerikkalainen eteläamerikkalaiseen eteläamerikkalaiselle eteläamerikkalaisen eteläamerikkalaisessa eteläamerikkalaisesta eteläamerikkalaiset eteläamerikkalaisia eteläamerikkalaisiin eteläamerikkalaisille eteläamerikkalaisilta eteläamerikkalaisina eteläamerikkalaista eteläamerikkalaisten) #syn(velkakriisi velkakriisiin velkakriisin velkakriisistä velkakriisiä) #syn(velkaantumisongelma velkaantumisongelman) #syn(kehittyessä kehittyessään kehittyi kehittyisi kehittyisikö kehittyisivät kehittyivät kehittymistä kehittymistäkin kehittymistään kehittymässä kehittymästä kehittymään kehittynee kehittyneeksi kehittyneelle kehittyneellä kehittyneemmillä kehittyneemmiltä kehittyneemmissä kehittyneemmälle kehittyneemmän kehittyneemmät kehittyneempi kehittyneempien kehittyneempikin kehittyneempiä kehittyneempänä kehittyneempää kehittyneempään kehittyneen kehittyneenä kehittyneeseen kehittyneessä kehittyneestä kehittyneet kehittyneiden kehittyneille kehittyneillä kehittyneiltä kehittyneimmille kehittyneimmistä kehittyneimmän kehittyneimmät kehittyneimpien kehittyneimpiin kehittyneimpiä kehittyneimpään kehittynein kehittyneintä kehittyneisiin kehittyneissä kehittyneissäkään kehittyneistä kehittyneitten kehittyneitä kehittynevät kehittynyt kehittynytkin kehittynyttä kehittyvien kehittyviin kehittyville kehittyvillä kehittyvine kehittyvissä kehittyvistä kehittyviä kehittyvä kehittyväksi kehittyvälle kehittyvällä kehittyvän kehittyvänkin kehittyvänä kehittyvässä kehittyvästä kehittyvät kehittyvää kehittyvään kehittyy kehittyä kehittyäkseen kehity kehitykään kehityttyä kehityttävä) #syn(ongelma ongelmaa ongelmaamme ongelmaan ongelmaansa ongelmahan ongelmakseen ongelmaksi ongelmalle ongelmalta ongelmamme ongelman ongelmana ongelmanaan ongelmanamme ongelmani ongelmanne ongelmansa ongelmassa ongelmasta ongelmastaan ongelmat ongelmia ongelmiaan ongelmiahan ongelmiakaan ongelmiakin ongelmiamme ongelmianne ongelmiemme ongelmien ongelmiensa ongelmiimme ongelmiin ongelmiinsa ongelmiksi ongelmilla ongelmille ongelmilta ongelmina ongelminaan ongelmineen ongelmissa ongelmissaan ongelmissakin ongelmista ongelmistaan ongelmistani ongelmitta ongelmaisesta ongelmaisia ongelmaista ongelmallinen ongelmallinenkin ongelmalliseen ongelmalliseksi ongelmallisella ongelmalliselta ongelmallisemmaksi ongelmallisempaa ongelmallisempana ongelmallisempi ongelmallisempia ongelmallisen ongelmallisena ongelmallisessa ongelmallisesta ongelmallisesti ongelmalliset ongelmallisia ongelmallisiakin ongelmallisiin ongelmallisiksi ongelmallisilla ongelmallisimmaksi ongelmallisimmasta ongelmallisimmillaan ongelmallisimmissa ongelmallisimpia ongelmallisimpiin ongelmallisin ongelmallisina ongelmallisinta ongelmallisissa ongelmallisista ongelmallista ongelmallisten ongelmaton ongelmatonta ongelmattomaksi ongelmattomampi ongelmattomana ongelmattomasti ongelmattomat ongelmattomia ongelmattomiksi ongelmattomin ongelmattominta) #syn(pyrityn pyrimmekin pyrimme pyrikään pyri pyrittävän pyrittäväksi pyrittävä pyrittäneen pyrittäisiin-

kin pyrittäisiin pyrin pyritte pyrittiin pyritty pyrittykään pyrittyään pyrittäessä pyrittäisi pyritä pyritäkään pyritään pyritäänhän pyritäänkin pyritäänkö pyrki pyrkien pyrkiessä pyrkiessämme pyrkiessään pyrkii pyrkiikin pyrkiikö pyrkikin pyrkikäämme pyrkikö pyrkimistä pyrkimällä pyrkimässä pyrkimättä pyrkimään pyrkinee pyrkineekin pyrkineelle pyrkineen pyrkineensä pyrkineessä pyrkineet pyrkineetkään pyrkineiden pyrkineiltä pyrkineistä pyrkineitä pyrkinevät pyrkinyt pyrkinytkin pyrkinytkään pyrkinyttä pyrkisi pyrkisikö pyrkisimme pyrkisin pyrkisivät pyrkivien pyrkiviin pyrkiville pyrkivillä pyrkiviltä pyrkivinä pyrkivissä pyrkivistä pyrkiviä pyrkivä pyrkiväksi pyrkivälle pyrkivällä pyrkivältä pyrkivän pyrkivänsä pyrkivänä pyrkivässä pyrkivästä pyrkivät pyrkivätkin pyrkivää pyrkivään pyrkiä pyrkiäkin pyrkiäkseen pyrkijä pyrkijällä pyrkijän pyrkijästä pyrkijät pyrkijää pyrkijöiden pyrkijöiksi pyrkijöille pyrkijöiltä pyrkijöistä pyrkijöitä pyrkiminen pyrkimiseen pyrkimisen pyrkimisessä pyrkimisestä pyrkimisestään pyrkimisiä pyrkimistä) #syn(ratkaise ratkaisee ratkaiseekin ratkaisemaan ratkaisemalla ratkaisemassa ratkaisematta ratkaisemiksi ratkaisemin ratkaisemista ratkaisemistaan ratkaisemme ratkaiseva ratkaisevaa ratkaisevaan ratkaisevaksi ratkaisevaksikin ratkaisevalla ratkaisevalle ratkaisevalta ratkaisevammaksi ratkaisevammassa ratkaisevammin ratkaisevampaa ratkaisevampi ratkaisevan ratkaisevana ratkaisevani ratkaisevansa ratkaisevassa ratkaisevasta ratkaisevasti ratkaisevastikin ratkaisevat ratkaisevia ratkaisevien ratkaiseviin ratkaiseviksi ratkaiseville ratkaisevimmaksi ratkaisevimmassa ratkaisevimmat ratkaisevimmin ratkaisevimpaan ratkaisevimpana ratkaisevimpiin ratkaisevin ratkaisevina ratkaisevinta ratkaisevista ratkaisi ratkaisimme ratkaisisi ratkaisisivat ratkaisivat ratkaiskaa ratkaiskoon ratkaisse ratkaissee ratkaisseemme ratkaisseen ratkaisseensa ratkaisseet ratkaisseista ratkaissut ratkaista ratkaistaan ratkaistaanko ratkaistaessa ratkaistaisiin ratkaistakaan ratkaistakseen ratkaistaksemme ratkaistane ratkaistaneen ratkaistava ratkaistavaa ratkaistavakseen ratkaistavaksi ratkaistavalta ratkaistavan ratkaistavana ratkaistavanaan ratkaistavat ratkaistavia ratkaistaviin ratkaistaviksi ratkaistavina ratkaistavissa ratkaistavista ratkaistessa ratkaistessaan ratkaistiin ratkaistu ratkaistua ratkaistuaan ratkaistuiksi ratkaistuja ratkaistuksi ratkaistun ratkaistuna ratkaistusta ratkaistut ratkaisija ratkaisijaan ratkaisijaksi ratkaisijalle ratkaisijan ratkaisijana ratkaisijat ratkaisijoiksi ratkaisijoina ratkaisijoista ratkaisijoita ratkaisijoitten ratkaisija ratkaisijaan ratkaisijaksi ratkaisijalle ratkaisijan ratkaisijana ratkaisijat ratkaisijoiksi ratkaisijoina ratkaisijoista ratkaisijoita ratkaisijoitten));