RIITTA ALKULA

Merkkijonoista suomen kielen sanoiksi

Suomen kielen morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja liittämisen vaikutukset tekstin tallennukseen ja hakuun

English summary

Tampereen yliopisto
Tampere 2000

Merkkijonoista suomen kielen sanoiksi

Acta Universitatis Tamperensis 763

AKATEEMINEN VÄITÖSKIRJA
Tampereen yliopisto, informaatiotutkimuksen laitos

Myynti

Tampereen yliopiston
julkaisujen myynti
PL 617
33101 Tampere

Puh. (03) 215 6055
Fax (03) 215 7150
taju@uta.fi
http://granum.uta.fi

Kannen suunnittelu
Juha Siro

Acta Universitatis Tamperensis 763
ISBN 951-44-4885-5
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 51
ISBN 951-44-4886-3
ISSN 1456-954X

Tampereen Yliopistopaino Oy Juvenes Print
Tampere 2000

TIIVISTELMÄ

Tutkimuksessa on selvitetty, miten suomen kielen morfologisten tulkinta-ohjelmien avulla voidaan ratkaista sellaisia tiedon tallennuksen ja haun ongelmia, jotka johtuvat suomen kielen erityispiirteistä.

Tutkimusta varten rakennettiin oma testausympäristönsä, jossa samasta tekstiaineistosta tuotettiin joukko erilaisia tietokantoja. Eri tietokantoja luotaessa sovellettiin usealla eri tavalla suomen kielen morfogisia tulkintaohjelmia, ja näitä tietokantoja sekä niistä tehtyjen tiedonhakujen tuloksia vertailtiin toisiinsa. Projekti oli siis luonteeltaan laboratorioympäristössä toteutettu evaluointitutkimus. Testauksissa käytetty hakujärjestelmä oli käänteishakemistoon perustuva, Boolen operaattoreita käyttävä BASIS. Vertaillut tutkimusympäristöt olivat seuraavat:

T1) **Perinteinen hakeminen**: hakijan katkaisemat hakusanat (kysely kohdistui taivutusmuotohakemistoon, joka sisälsi dokumenttien sanat sellaisinaan, taivutusmuodoissaan)

T2) **Automaattinen katkaisu**: perusmuotoisten hakusanojen syöttäminen taivutusvartaloita tuottaville ohjelmille, kysely vartaloilla (kysely kohdistui taivutusmuotohakemistoon)

T3) **Seulonta**: automaattinen taivutusvartaloiden tuottaminen, kysely vartaloilla sekä tulosten seulonta perusmuotoon palauttavalla ohjelmalla (kysely kohdistui taivutusmuotohakemistoon)

T4) **Perusmuotojen** ja yhdyssanojen alkuosien hakeminen (kysely kohdistui perusmuotohakemistoon, jossa morfologisen tulkinta-ohjelman tunnistamat sanat olivat perusmuodossa, tunnistamatta jääneet sanat taivutusmuotoisina)

T5) **Perusmuotojen ja** yhdyssanan kaikkien **osien** hakeminen (kysely kohdistui ositettuun perusmuotohakemistoon, jonne perusmuotojen lisäksi on tallennettu yhdyssanoista kaikki niiden osat sekä näiden osien yhdistelmät)

T6) Perusmuotojen ja yhdyssanojen osien hakeminen; jos kysely perusmuodoilla ei tuottanut tulosta, haettiin vartalo-ohjelmien tuottamilla vartaloilla taivutusmuotohakemistosta (kysely kohdistui kaksoishakemistoon, eli perusmuotohakemiston ja taivutusmuotohakemiston yhdistelmään)

Kun hakemistoon tallennettavat sanat perusmuotoistettiin, perusmuotohakemistoon tuli vähemmän sanoja kuin taivutusmuotohakemistoon. Osoit-

teita taas tuli enemmän kuin taivutusmuotohakemistoon, mikä johtui monitulkintaisista sananmuodoista sekä yhdyssanan osista, mikäli myös nämä osat tallennettiin hakemistoon. Sanojen vähentymisen ja osoitteiden lisäyksen lopputuloksena kuitenkin oli, että perusmuotohakemisto vei vähemmän muistitilaa kuin taivutusmuotohakemisto. Tämä päti myös ositetussa perusmuotohakemistossa eli kun hakemistoon tallennettiin perusmuotojen lisäksi myös yhdyssanojen osat ja niiden yhdistelmät.

Perusmuotohakemistosta saatiin tarkempia tuloksia kuin taivutusmuotohakemistosta: kun molemmissa käytettiin taivutusvartalo-ohjelman tuottamia, täsmälleen samalla tavalla katkaistuja hakusanoja, perusmuotohakemistosta saatujen tulosjoukkojen tarkkuus oli parempi kuin samanlaisella kyselyllä taivutusmuotohakemistosta saatujen tulosjoukkojen tarkkuus.

Toisaalta perusmuotohakemistoon tehdyissä kyselyissä ei kannata käyttää pelkkiä hakusanan perusmuotoja. Kun perusmuotohakemistosta haettiin antamalla muuten samat hakusanat kuin taivutusmuotohakemistosta haettaessa, mutta jättämällä ne katkaisematta, saanti romahti. Kun kyselyyn lisättiin hakusanan perusmuotojen lisäksi sen johdokset tai hakusanan sisältävät yhdyssanat, tulosjoukon saanti nousi useampia prosenttiyksikköjä kuin tarkkuus samalla laski. Perusmuotohakemistoista haettaessa on siis muistettava ottaa myös johdokset ja yhdyssanat huomioon.

Korkeimmat saantiarvot saatiin ositetusta perusmuotohakemistosta (T5) ja perusmuotohakemistosta (T4) Erot toisiin tutkimusympäristöihin olivat systemaattisia, mutta yleensä eivät tilastollisesti merkitseviä. Samalla näiden kahden ympäristön tarkkuusarvot olivat paremmat kuin haettaessa hakijan katkaisemilla hakusanoilla taivutusmuotohakemistosta. Parhaat tarkkuusarvot saatiin seulottaessa (T3) ja perusmuotohakemistosta (T4). Seulonnalla saatujen hakutulosten saanti oli kuitenkin kaikkein huonoin; sen saannin ero verrattuna toisten menetelmien saantiin oli myös tilastollisesti merkitsevä.

Perusmuotoistamisen riskinä on, että tulkintaohjelmille tuntemattomat sanat tulkitaan väärin, jolloin hakemistoon päätyy vääriä sanoja. Tutkimuksessa kokeiltiin muutamia yksinkertaisia korjausmenetelmiä, joilla tällaiset väärät tulkinnat voidaan hakuvaiheessa kiertää ja siten varmistaa dokumenttien löytyminen tai parantaa hakutulosten tarkkuutta.

SUMMARY

This thesis deals with linguistic processing and retrieval techniques in Finnish fulltext databases, and especially the special problems due to the characteristics of the Finnish language.

In the research project, natural language analysis modules for Finnish were incorporated into the commercial BASIS information retrieval system, which is based on inverted files and Boolean searching. Several test databases were produced, each using one or two Finnish morphological analysis modules. Thus the work has been carried out within the traditional laboratory environment paradigm, presenting various test parameters and their evaluation. The test databases were as follows:

T1) **Traditional** database: the inflected word forms were stored as such in the index and the searcher truncated search terms manually

T2) **Automatic stemming**: traditional index, but a language module was used to stem the search terms (linguistic stemming using the base form as input)

T3) **Sifting the results of automatic stemming**: As above, but the character strings that matched stemmed strings were further analyzed. Only strings that were variants of the original search term were accepted; false drops having only the beginning of the word identical with the search string were rejected.

T4) **Base word forms**: The inflected word forms were reduced to their base form before storing the words in the index. The searcher used the base form as a search term. When compound words were searched, the basic word form was fed into the automatic stemming module (as in T2 above, but using a different index)

T5) **Base word forms + split compounds**: The inflected word forms were reduced to their basic form and compound words were broken down into their components before storing the words in the index. The basic word form was used as a search term. Compound words were searched by adding a special "compound wild card" symbol.

T6) **Double index**: consisted of both the inflected index and the base word form index. (The purpose was to test how to recover from errors made in base word form generation.)

When word forms were normalized to their base form, the memory size of the base form index file was smaller than the size of the inflected word form index. Even when the compounds were broken down into their components, the base form index was smaller than the inflected form index.

When automatic stemming was used in the different databases, the precision figures were higher when the searches were conducted in the base form index compared to searching with the same search stems in the inflected word form index. This indicates that a base form index is inherently more precise than an inflected word form index.

When searching in base form index, it is not practical to use only the base form of the original search term. When a simple query consisting of only the original search term in its base form is conducted, the recall ratio will collapse, compared to the corresponding search with truncated search terms in the inflected word form index. The recall ratio becomes better when such a simple query is augmented with the derivatives and/or compound words related to the search term. Although the precision ratio simultaneously will decrease, the loss in precision will be smaller than the profit in recall. Consequently, derivatives and compounds are of importance when searching in base word form index.

The best recall ratio was achieved in the index containing the base word forms (T4) and the one containing the base forms as well as the components of compound words (T5). Differences from other databases were systematic, but usually not statistically significant.

The best precision ratio was achieved when the search terms were sifted (T3) and when using the base word form index (T4). Sifting, however, obtained the poorest recall ratio, and the differences from other databases were also statistically significant.

Problems in using natural language modules were also investigated. One problem is that language modules may provide a false reading of an inflected word form. This may result in information loss (the correct word form will not be in the index) or information overflow (all possible readings of homonymic word forms must be stored to index, which increases the amount of index terms). Some simple methods for transforming improper search terms to appropriate base forms were tested.

ALKUSANAT

Tämä opinnäyte on syntynyt Suomenkielisten tekstitietokantojen tallennus- ja hakutekniikat (FULLTEXT) -projektin tulosten pohjalta. Projektin tuloksia on aiemmin raportoitu muun muassa Valtion teknillisen tutkimuskeskuksen (VTT) Julkaisuja-sarjassa (Alkula & Honkela 1992). Tässä väitöskirjassa perehdytään tutkimuksen tiettyihin osa-alueisiin aikaisempia julkaisuja laajemmin ja syvällisemmin.

FULLTEXT-projektin pääasiallinen rahoittaja oli Teknologian kehittämiskeskus (TEKES). Muita rahoittajia olivat Valtion tietokonekeskus (VTKK, nykyisin TietoEnator, Tietopalvelut), KTA-Papyrus Oy ja VTT:n informaatiopalvelulaitos (nykyisin VTT Tietopalvelu). Projektissa oli kaksi testausympäristöä: BASIS-hakujärjestelmä ja kokeellinen APL-Minttu-hakujärjestelmä. Käsillä olevassa tutkielmassa selostetaan BASIS-hakujärjestelmän testauksia ja testituloksia, joiden toteuttaminen oli FULLTEXT-projektissa tämän kirjoittajan vastuulla. APL-Minttu-hakujärjestelmän testauksista vastasi VTT:n tietojenkäsittelytekniikan laboratorion tutkija Timo Honkela (nykyisin Taideteollisessa korkeakoulussa). Tässä opinnäytetyössä nimitetään yksinkertaisuuden vuoksi projektin pelkkää BASIS-osuutta FULLTEXT-projektiksi. Mikäli tekstissä selostetaan APL-Minttu-ympäristössä tehtyjä testauksia, siitä mainitaan erikseen. Tarkemmin APL-Minttu-testaukset on kuvattu VTT:n julkaisussa (Alkula & Honkela 1992).

BASIS-hakujärjestelmän testauksiin osallistuivat ja myötävaikuttivat seuraavat henkilöt: Markku Ylinen Aamulehdestä (Almamedia) järjesti Aamulehden sähköisen arkiston artikkeleita projektin tutkimustietokannaksi. Helsingin Sanomien Tarja Hjorthin ja Kaarina Nazarenkon avustuksella projekti sai käyttöönsä Sanomien arkistossa muistiinmerkittyjä toimittajien hakupyyntöjä. Testikyselyjä arvioivat VTT:n informaatiopalvelulaitoksen informaatikot Jaakko Anttila, Kari Martiskainen ja Irma Salovaara. KTA:n Markku Kuokkala ja Raili Salminen liittivät suomen kielen tulkintaohjelmat BASIS-hakujärjestelmään. Lisäksi Raili Salminen vastasi testitietokantojen tuottamiseen ja seurantatietojen keruuseen tarvittavista tietokoneajoista. Testikyselyjen suorittamisessa ja tulosjoukkojen käsittelyssä avusti Pirjo Valpas. Hakujen tuloksena saatujen sanomalehtiartikkeleiden käyttökelpoisuutta arvioivat Länsiväylälehden toimittajat Tarja Heinivaho, Klaus Nurmi ja Tuija Tuominen.

Opinnäytetyön ohjaajana on ollut Kalervo Järvelin Tampereen yliopiston Informaatiotutkimuksen laitokselta. Tämän työn käsikirjoitusta ovat eri vaiheissa kommentoineet Eero Sormunen, Heikki Keskustalo ja Jaana Kekäläinen Informaatiotutkimuksen laitokselta sekä Timo Honkela.

Niiden monien henkilöiden joukosta, joilta olen saanut apua tämän tutkimuksen toteuttamisessa ja raportoinnissa, haluan erityisesti mainita Kalervo Järvelinin, Eero Sormusen ja Markku Ylisen, joiden avulla projekti saatiin käyntiin ja aineisto hankituksi ja jotka sittemmin ovat kommentoineet tutkimusta ja sen tuloksia eri vaiheissaan. Lämpimät kiitokseni myös muille tutkimukseen osallistuneille.

Suurimmat kiitokset perheelleni siitä, että elämä ei ole ollut pelkkää työtä ja väitöskirjaa. Kiitokset Jormalle kärsivällisyydestä, vaikka opinnäytteen parissa puurtamiselleni ei vuosiin näyttänytkään tulevan loppua, ja Maaritille sen muistuttamisesta, että maailmassa on niin monta ihmeellistä asiaa.

Espoossa 3.4.2000 Riitta Alkula

Joh. 1:1

SISÄLLYSLUETTELO

1 JOHDANTO 15
1.1 Tutkimuksen lähtökohdat 15
1.2 Suomen kielen ominaisuudet tiedonhaun kannalta 18
1.3 Tutkimusongelman alustava määrittely 20
1.4 FULLTEXT-projekti 21

2 KÄSITTEIDEN MÄÄRITTELY 23
2.1 Tiedonhaun käsitteet 23
2.2 Kielitieteen käsitteet 30
2.3 Tutkimuksen erityiskäsitteet 32

3 TIEDONHAUN TULOKSELLISUUDEN MITTAAMINEN 36
3.1 Relevanssin määrittely 36
3.2 Saannin ja tarkkuuden laskeminen 39

4 KIELITIETEEN HYÖDYNTÄMINEN TIEDONHAKUTUTKIMUKSESSA 42
4.1 Kielen osajärjestelmät 42
4.2 Fonologia ja fonetiikka 44
4.3 Morfologia 48
4.3.1 Keinoja sananmuotojen yhdistämiseksi 49
4.3.2 Englannin karsinta-algoritmien kritiikki - onko karsinnalla väliä? 51
4.3.3 Karsinta-algoritmien edelleenkehittäminen 56
4.3.4 Muiden kielten karsinta-algoritmit 62
4.4 Leksikko 63
4.5 Syntaksi 64
4.5.1 Yleistä 64
4.5.2 Lausekerakenteen automaattisen tunnistaminen 67
4.5.3 Boolen operaattorien päätteleminen hakupyynnöstä 71
4.6 Semantiikka 72
4.7 Teksti ja diskurssi 76
4.7.1 Tilanteinen merkitys ja diskurssianalyysi 76
4.7.2 Tekstin sidoksisuus 78

5 TUTKIMUKSESSA KÄYTETYT OHJELMISTOT 81
5.1 BASIS-hakujärjestelmä 81
5.2 Suomen kielen morfologiset tulkintaohjelmat 84
5.2.1 Taivutusvartaloita tuottavat ohjelmat 85
5.2.2 Perusmuoto-ohjelmat 86

6 TUTKIMUSONGELMA JA TUTKIMUSHYPOTEESIT 90
6.1 Hakemistojen koko 91
6.2 Taivutusmuotohakemiston hakuominaisuudet 93
6.3 Perusmuotohakemiston hakuominaisuudet 94
6.4 Ositetun perusmuotohakemiston hakuominaisuudet 96
6.5 Vakio- ja ongelmakyselyt 97

7 TUTKIMUKSEN TOTEUTUS 99
7.1 Tekstiaineiston hankinta 99
7.2 Yhdyssanan osiin jakamisen ja tallentamisen vaihtoehdot 100
7.3 Tuotetut hakemistot 108
7.3.1 Toteutuksen periaatteet 108
7.3.2 Taivutusmuotohakemisto 110
7.3.3 Perusmuotohakemistot 111
7.3.4 Kaksoishakemisto 112
7.4 Tutkimusympäristöt 113
7.4.1 Toteutuksen periaatteet 113
7.4.2 Perinteinen hakeminen 115
7.4.3 Automaattinen katkaisu 116
7.4.4 Seulonta 117
7.4.5 Perusmuotojen ja yhdyssanojen alkuosien hakeminen 120
7.4.6 Perusmuotojen ja yhdyssanojen kaikkien osien hakeminen 120
7.4.7 Hakeminen kaksoishakemistosta 123
7.5 Syötteen tarkistus- ja virheenkorjausmenetelmien yleisperiaatteet 123
7.6 Testikyselyjen laatiminen 126
7.6.1 Hakupyyntöjen keräys ja valinta 126
7.6.2 Hakusanojen valinta 128
7.6.3 Kyselytyyppien muodostaminen eri tutkimusympäristöjen vertailua varten 130
7.7 Hakutulosten relevanssiarviot 138
7.8 Saannin ja tarkkuuden laskeminen 141
7.9 Otantaperiaatteet 142
7.10 Tilastollisen merkitsevyyden laskeminen ja eri vaihtoehtojen välisen eron merkittävyys 144

8 PERUSMUOTOISTAMISEN VAIKUTUS HAKEMISTOSANOIHIN 149
8.1 Tallennettavan tekstin käsittely 149
8.2 Merkkijonojen määrä hakemistossa 151
8.3 Osoitteiden määrä hakemistossa 153

8.4 Hakemiston muistitila 155
8.5 Twol-ohjelmiston testaus 159
8.6 Hakemisto käyttäjän silmin 161

9 PERUSMUOTOISET HAKUSANAT VAKIOKYSELYISSÄ 168
9.1 Taivutusmuotohakemistoon perustuva tutkimusympäristö 170
9.2 Hakijan katkaisemien ja automaattisesti katkaistujen hakusanojen vertailu 170
9.2.1 Finstemsin tuottamilla vartaloilla hakeminen 171
9.2.2 Finstems- ja Hahmotin-ohjelmien väliset erot 177
9.3 Hakijan katkaisemien ja seulottujen hakusanojen vertailu 179
9.3.1 Perusjoukko 179
9.3.2 Johdososajoukko 181
9.3.3 Yhdyssanaosajoukko 182
9.3.4 Seulonnan toteutustavasta 185
9.4 Taivutusmuoto- ja perusmuotohakemistosta saatujen hakutulosten vertailu 186
9.4.1 Perusjoukko 187
9.4.2 Johdososajoukko 191
9.4.3 Yhdyssanaosajoukko 195
9.5 Taivutusmuotohakemistosta ja ositetusta perusmuotohakemistosta saatujen hakutulosten vertailu 201
9.5.1 Perusjoukko 201
9.5.2 Johdososajoukko 205
9.5.3 Yhdyssanaosajoukko 209
9.6 Eri tutkimusympäristöjen vertailu keskenään 215
9.6.1. Perusjoukko 215
9.6.2. Johdososajoukko 216
9.6.3. Yhdyssanaosajoukko 217
9.6.4. Yhteenveto 219

10 ONGELMAKYSELYT 224
10.1 Ongelmakyselyjen tyypit 226
10.2 Automaattisesti katkaistujen hakusanojen tarkistus taivutusmuotohakemistosta haettaessa 227
10.2.1 Aidosti tuntematon tai sananmuotohomografikseen tulkittu perusmuoto 227
10.2.2 Taivutusmuodossa annettu hakusana 229
10.2.3 Yhteenveto 231
10.3 Syötteen tarkistaminen perusmuotohakemistoista haettaessa 231
10.3.1 Aidosti tuntematon perusmuoto 231
10.3.2 Sananmuotohomografikseen tulkittu perusmuoto 234
10.3.3 Taivutusmuodossa annettu hakusana 236

10.4 Syötteen tarkistaminen kaksoishakemistoista haettaessa 238

11 TULOSTEN TARKASTELU 239
11.1 Koejärjestelyt 239
11.2 Hypoteesien toteutuminen 241
Hypoteesi 1 241
Hypoteesi 2 243
Hypoteesi 3 245
Hypoteesi 4 247
Hypoteesi 5 249
Hypoteesi 6 253
Hypoteesi 7 253
11.3 Homografiasta johtuvat ongelmat 256
11.4 Sulkusanalista 259
11.5 Syötteen tarkistus- ja virheenkorjausmenetelmät 261
11.5.1 Automaattisesti katkaistujen hakusanojen tarkistus 261
11.5.2 Syötteen tarkistus hakutulosten seulonnan yhteydessä 263
11.5.3 Syötteen tarkistaminen perusmuotohakemistosta haettaessa 263
11.5.4 Syötteen tarkistaminen kaksoishakemistosta haettaessa 265
11.6 Ongelmasanojen yleisyys 267
11.7 Läheisyysoperaattorin hyöty sanaliittoja haettaessa 268

12 JOHTOPÄÄTÖKSET 270

LÄHDELUETTELO 278
Painetut lähteet 278
Painamattomat lähteet 293

LIITTEET

1 JOHDANTO

1.1 Tutkimuksen lähtökohdat

Tekstitietokantoihin tallennetut tekstit eivät sisällä mitä tahansa satunnaisia merkkijonoja, vaan jonkin tietyn kielen ilmauksia: virkkeitä, sanoja, sanaliittoja ja niin edelleen. Näin ollen tuntuisi järkevältä, että luonnollisen kielen ominaisuuksia hyödynnettäisiin tekstejä tallennettaessa ja haettaessa sen sijaan, että käsiteltäisiin pelkästään merkkijonoja. Monet tekstin ominaisuudethan riippuvat siitä, millä kielellä se on kirjoitettu.

Kaupallisten hakujärjestelmien kehitys on tähän asti käytännössä tapahtunut englannin kielen ehdoilla. Englanninkielisten sanojen käsittelyyn ovat riittäneet melko yksinkertaiset menetelmät, joten hakujärjestelmien kehittäjät ovat katsoneet, että luonnollisen kielen ominaisuuksia hyödyntäviä menetelmiä ei tarvita, koska niistä saatava hyöty ei välttämättä ylitä kehittämis- ja käyttöönottokustannuksia. Sitä paitsi kaupallisia tiedonhakujärjestelmiä ei ole kehitetty erityisen systemaattisesti: toimintoja on lisätty, muutettu ja poistettu käyttäjäkunnan toivomusten mukaan tai kopioimalla kilpailevissa hakujärjestelmissä toteutettuja ideoita (Tenopir & Ro 1990, s. 49). Hakujärjestelmien käytössä ilmenneiden ongelmien perimmäisiä syitä ja ratkaisuja ei välttämättä ole pohdittu - etenkin jos tilanteen korjaaminen vaatisi hakujärjestelmän tai tietokantojen rakenteen perusteellista muutosta - vaan on kehitetty toiminto, jonka avulla ongelma voidaan kiertää tai ratkaista edes osittain.

Tilanne on kuitenkin muuttunut, kun tekstin tallennus- ja hakumenetelmille asetetut vaatimukset ovat kasvaneet. Kun tekstit tuotetaan ja muokataan suoraan tietokoneella, on suurtenkin tekstiaineistojen tallentaminen ja arkistoiminen elektronisesti tullut taloudellisesti yhä edullisemmaksi. Samalla kun elektronisten tekstitietokantojen ja tekstiarkistojen sekä yksittäisten dokumenttien koko on kasvanut, perinteiset tallennus- ja hakumenetelmät ovat alkaneet jäädä tehottomiksi. Toisaalta tekstitietokantojen käyttäjäkunta ei enää ole vain tiedonhakuun erikoistuneita ammattilaisia, joten samalla tekstitietojärjestelmien käytön halutaan olevan aiempaa yksinkertaisempaa. Tämä havainnollistuu erityisesti Internetissä, jossa tätä kirjoitettaessa on jo yli miljardi WWW-sivua (Inktomi 2000). Internet-verkostossa periaatteessa kuka tahansa voi poimia katseltavakseen dokumentin mistä päin maailmaa tahansa. Jos tiedontarvitsija ei tiedä häntä kiinnostavan dokumentin täsmällistä Internet-osoitetta, hän voi käyttää

hyväkseen erilaisia hakupalveluja (esimerkiksi eXcite, AltaVista), joiden avulla kuka tahansa pääsee - tai joutuu - etsimään haluamiaan dokumentteja sanahaun avulla.

Tekstidokumenttien tallennuksessa ja haussa siis tarvitaan yhä tehokkaampia ja toisaalta helpompia menetelmiä, jotta suurista tekstimassoista saadaan poimituksi vain tarpeelliset dokumentit. Suurissa tietokannoissa esimerkiksi hakutuloksen tarkkuus on huomattavasti kriittisempi tekijä kuin pienissä tietokannoissa, koska epätarkasti muotoiltu kysely voi palauttaa enemmän dokumentteja kuin käyttäjä on halukas käymään läpi (Blair 1990; Ledwith 1992; Sormunen 1994 ja 2000).

1990-luvun alkuvuosiin asti tiedon tallennuksen ja haun ongelmia oli suhteellisen vähän yritetty ratkaista kielitieteen menetelmien tai teorioiden avulla (Warner 1991). Vaikka molempien tutkimusalueiden kohteena ovatkin luonnollisella kielellä esitetyt tekstit, ne kuitenkin painottavat eri seikkoja. Tiedon tallennuksen ja haun tutkimuksen päätavoite on kehittää käsitteitä, menetelmiä ja järjestelmiä, joiden avulla kaikki tieto, missä tahansa muodossa tai paikassa se onkin, saadaan vaivattomasti sitä tarvitsevan ulottuville ja siinä muodossa, että tiedontarvitsijan on mahdollisimman helppo tämä tieto omaksua (Järvelin & Niemi 1990; Järvelin 1995). Kielitiede puolestaan on kiinnostunut tekstiin sisältyvän luonnollisen kielen rakenteesta ja käyttäytymisestä sinänsä (eli haluaa selittää, miten luonnollinen kieli toimii), millä on tiedon tallennuksen ja haun kannalta vain välinearvo. Luonnollisen kielen rakenne ja muut ominaisuudet kiinnostavat tiedon tallennuksen ja haun tutkimusta siksi, että niiden avulla pyritään pääsemään käsiksi haluttuun tietoon tehokkaammin.

Ensimmäinen kausi, jolloin luonnollisen kielen käsittelyä yritettiin toden teholla hyödyntää tiedonhaun tutkimuksessa, kesti 1950-luvun puolivälistä 1960-luvun puoleenväliin. Tuonaikaiset tutkimusasetelmat olivat kuitenkin ylioptimistisia niin kielitieteen kuin tietojenkäsittelytekniikankin suhteen: luonnollisen kielen ominaisuudet eivät olleet niin helposti automatisoitavissa kuin oli kuviteltu. Tutkimusalue joutui vakavaan kriisiin, kun tietokoneista ei ollutkaan vastaavaa hyötyä alueilla, joissa niitä innokkaimmin yritettiin hyödyntää eli automaattisessa indeksoinnissa ja tiedonhaussa. Vaikka tietokoneilla pystyttiin nopeasti käsittelemään tekstiin sisältyviä sananmuotoja eli käytännössä merkkijonoja, näiden merkityssisältö jäi tavoittamatta. (Sparck Jones & Kay 1973)

Myöhemmin 1960- ja 1970-luvuilla vallinneet kielioppiteoriat, kuten transformaatiokielioppi, olivat hankalasti tiedonhaun käytäntöihin sovellettavissa. Kielitieteessä keskityttiin hienosyiseen lausetason analyysiin. Informaatiotutkimus taas oli kielitieteilijöiden näkökulmasta liian pragmaattista ollakseen tieteellisesti kiinnostavaa. Perinteiset kielitieteilijät rajautuivat puhtaaseen tieteeseen ja vieroksuivat sellaista, mikä vivahti insinööritieteeseen tai käytännön sovelluksiin. Samasta syystä kielitieteilijät suhtautuivat varauksellisesti myös tietokonelingvistiikkaan. Tietokonelingvistiikka on monitieteinen alue, jonka perustavoitteena on kiel(t)en rakenteen ja toiminnan jäljittely tietokoneen avulla tavalla, joka on lingvistisesti perusteltu, formaalisesti täysin ymmärretty ja tietokoneohjelmana toteutettu (Karlsson 1994). Spark Jones ja Kay (1973) kuitenkin ennustivat, että kielitieteessä juuri tietokonelingvistiikka on todennäköisimmin se tutkimusalue, josta on odotettavissa eniten hyötyä tiedon tallennuksen ja haun tutkimukselle.

1990-luvulla tiedonhakututkimuksessa kuitenkin virisi uusi mielenkiinto kielitieteellisiä menetelmiä kohtaan, mikä oli nimenomaan tietokonelingvistiikan kehityksen ansiosta. Sen tuloksena on syntynyt kielioppiteorioita ja menetelmiä, jotka ovat aiempaa paremmin tiedonhakuun sovellettavissa. Vuonna 1993 tapahtui varsinainen läpimurto, kun tiedonhaku tavallisella englannin kielellä (plain English searching) tuli mahdolliseksi useissa suurissa kaupallisissa tiedonhakujärjestelmissä (WESTLAW, LEXIS-NEXIS, DIALOG). Aiemmin tällainen oli ollut mahdollista vain kokeellisissa hakujärjestelmissä (Hattery 1993). Osoitus kasvaneesta mielenkiinnosta on myös se, että vuonna 1996 TREC[1]-hankkeessa perustettiin oma ryhmänsä tutkimushankkeille, jotka hyödynsivät luonnollisen kielen käsittelyä (Strzalkowski et al. 1997).

Viime vuosina tiedonhakututkimuksen kohteina olleet muun muassa seuraavat alueet (Pirkola 1999): monikielinen tiedonhaku, vaihtoehtoisten hakuavainten tuottaminen, hakuavainten muokkaaminen morfologisin keinoin, hakuavainten monimerkityksisyyden poistaminen (yksiselitteistäminen), ellipsien ja anaforien avaaminen, lause(ke)rakenteiden automaattinen tunnistaminen sekä puheeseen

[1] TREC, Text Retrieval Conference, on hanke, jossa rakennetaan suurta, kansainvälistä, yleisessä käytössä olevaa tiedonhakujen tutkimusympäristöä. Siihen kuuluu joukko erilaisia tekstikokoelmia, joukko hakupyyntöjä sekä kullekin hakupyynnölle arviot siitä, mitkä kokoelman dokumentit vastaavat ko. hakupyynnön tiedontarvetta. Standardoituja aineistoja käyttämällä pyritään helpottamaan eri tiedonhakumenetelmien välistä vertailua. <URL: http//:trec.nist.gov/>

perustuva tiedonhaku. Tulokset ovat olleet vaihtelevia: joissain tutkimuksissa kielitieteellisistä menetelmistä on todettu olevan hyötyä (esimerkiksi Krovetz 1993; Hull 1996), joissain taas niistä ei ole näyttänyt olevan erityisempää hyötyä, vaan melkeinpä haittaa (Arampatzis et al. 1998). Sopivien sovellusalueiden löytäminen kielitieteellisille menetelmille on siis edelleenkin tiedonhaun tutkijoiden haasteena.

Tiedonhakututkimuksen valtavirta suuntautuu englanninkielisen tekstin ongelmien tutkimiseen ja ratkaisemiseen. Selvää kuitenkin on, että suomalaisiinkin hakujärjestelmiin olisi tarpeen saada menetelmiä, jotka tukevat suomen kielellä tehtävää tiedonhakua. Tämä kuitenkin edellyttää, että tarvittava suomen kielen tulkintatekniikka on käytettävissä. Tämän tutkimuksen aikana tällainen tekniikka oli morfologian osalta olemassa: suomen kielen automaattista tulkintaa tutkineet kaksi tutkimusryhmää ovat 1980-luvulta alkaen kehittäneet tulkintaohjelmia, jotka ovat olleet kansainvälistä huipputasoa (erityisesti Koskenniemi 1983). Morfologisia tulkintaohjelmia on sovellettu muun muassa suomenkielisen tekstin tavuttamiseen ja oikolukuun (Karlsson 1985, Jäppinen et al. 1983). 1990-luvulla morfologiset tulkintaohjelmat ovat molemmissa suomalaisissa tutkimusryhmissä olleet vakiintunutta perustekniikkaa, jonka päälle on rakennettu muun muassa lauseenjäsennysohjelmia (enemmän aiheesta luvussa 4).

1.2 Suomen kielen ominaisuudet tiedonhaun kannalta

Tiedonhakujärjestelmiin on aikojen myötä kehitetty erilaisia menetelmiä, joilla luonnollisen kielen vaihtelua pyritään hallitsemaan. Tällaisten menetelmien teho kuitenkin riippuu kohteena olevan kielen ominaisuuksista. Suomen kielen erikoispiirteistä seuraa tiedon tallennuksessa ja haussa ongelmia, joita nykyisissä tiedonhakujärjestelmissä ei voida tyydyttävästi ratkaista. Tämä johtuu siitä, että hakujärjestelmät on kehitetty ensisijaisesti englanninkielisen tekstin ominaisuuksien perusteella.

Suomen kielen nomineilla on 15 sijamuotoa. Sijapäätteiden lisäksi erilaisia nominien sananmuotoja muodostetaan tunnusten, johtimien ja liitteiden sekä näiden yhdistelmien avulla (koulu+i+ssa+mme+kin). Yhdestä verbistä taas voi muodostaa 136 persoonamuotoa. Kun otetaan huomioon kaikki teoreettisesti mahdolliset yhdistelmät, yhdellä substantiivilla on noin 2 000, adjektiivilla 6 000 ja verbillä 12 000 taivutusmuotoa. Näissä luvuissa ei ole vielä mukana

johdoksia. Johdosten huomioonottaminen kasvattaa edellämainitut luvut noin kymmenkertaisiksi. (Koskenniemi 1985a)

Perinteisissä hakujärjestelmissä tallennetaan dokumenteissa esiintyneet sanat hakemistoon sellaisinaan, taivutusmuotoineen kaikkineen. Kukin sanan esiintymä on oma hakemistosanansa, vaikka ne eroaisivat toisistaan vain yhden kirjaimen verran (esimerkiksi tytöllä : tytöillä). Vaikka suuri osa teoreettisesti mahdollisista sananmuodoista ei esiinnykään hakujärjestelmään tallennettavien dokumenttien tekstissä, yksittäisen sanan eri esiintymiä löytyy hakemistosta silti paljon.

Toinen suomen kielelle tyypillinen ominaisuus on yhdyssanojen runsaus. Esimerkiksi Nykysuomen sanakirjassa yhdyssanoja on noin 130 000 kappaletta eli yhdyssanojen osuus koko Nykysuomen sanakirjan sanamäärästä on suunnilleen kaksi kolmasosaa (Saukkonen 1973). Suomen kielessä yhdyssanan muodostavat perussanat kirjoitetaan yhteen (esimerkiksi ydinvoimalaitos), kun ne englannissa kirjoitetaan erillisinä sanoina (nuclear power plant). Suomenkielisestä tekstistä haettaessa ongelmana on, miten yhdyssanojen keski- ja loppuosat saataisiin löydetyksi, koska niitä on hankala hakea perinteisestä hakujärjestelmästä. Englanninkielisen hakijan ongelmana taas on, miten yhdyssanojen eri osat saadaan hakuvaiheessa kytkettyä niin tiukasti yhteen, että muut kuin yhdyssanaesiintymät karsiutuisivat pois.

Tekstitiedonhaun tulosten voidaan olettaa paranevan, kun käytetään tallennus- ja hakumenetelmiä, jotka ottavat kunkin kielen erityisominaisuudet huomioon. Kun hakusanat ovat suomen kielen sanoja, hakijalle on helpotus, että hänen itsensä ei tarvitse esimerkiksi pohtia suomen kielen sanojen taipumista. Suomenkielisen tekstin sisältämät eri sananmuodot on myös jollakin tavoin normalisoitava silloin, kun halutaan soveltaa edistyneitä tallennusmenetelmiä, joita on kokeiltu muissa, morfologisesti yksinkertaisemmissa kielessä. Tällaisia ovat esimerkiksi hypertekstilinkkien ja semanttisten verkkojen automaattinen tai puoliautomaattinen tuottaminen.

Suomen kielen (morfologisten) tulkintaohjelmien avulla voidaan toteuttaa hakujärjestelmä, joka muun muassa ottaa huomioon sanojen taipumisen ja huolehtii automaattisesti siitä, että hakusanan kaikki erilaiset esiintymät (sananmuodot) haetaan. Lisäksi tietyt tulkintaohjelmat pystyvät jakamaan yhdyssanat osiinsa.

Muutamassa suomalaisessa tutkimuksessa on esitetty erilaisia ratkaisuja, miten suomen kielen morfologisia tulkintaohjelmia voitaisiin liittää tiedonhakujärjestelmiin. Näissä tutkimuksissa asiaa on kuitenkin käsitelty periaatteellisesti, kokeilematta näiden ohjelmistojen toimintaa todellisessa hakujärjestelmässä (Nurminen 1986; Hjorth 1987b), tai sitten testauksissa käytetty tutkimusaineisto on ollut suhteellisen pieni (Niemistö 1988). Laajimmin suomenkielisen perusmuotohakemiston rakentamisen ongelmia on tähän mennessä tutkinut Keskustalo (1994).

Käytännössä suomen kielen tulkintaohjelmistoja on 1980-luvun lopusta hyödynnetty Aamulehden ja Satakunnan Kansan BASIS-ohjelmistolla toteutetuissa tekstiarkistoissa sekä Keskisuomalaisen arkistossa. Myös Helsingin Sanomat perusti vuonna 1992 vastaavanlaisen arkiston, jossa hakujärjestelmänä on BRS. Tulkintaohjelmien vaikutuksesta näiden lehtiarkistojen ominaisuuksiin ei kuitenkaan ole saatavilla julkisia vertailutietoja.

1.3 Tutkimusongelman alustava määrittely

Käsillä olevan opinnäytteen tutkimusongelma määritellään suppeasti ja yleisellä tasolla seuraavasti:

> Jos yksi tai useampia suomen kielen morfologisia tulkintaohjelmia liitetään osaksi tiedonhakujärjestelmää, niin voidaanko näiden tulkintaohjelmien avulla parantaa hakujärjestelmän hakemiston ominaisuuksia sekä nostaa hakutulosten saanti- ja tarkkuusarvoja?

Tutkimusongelma voidaan tarkentaa muun muassa seuraaviin osaongelmiin:

1) Jos dokumentissa esiintyneet sananmuodot perusmuotoistetaan ennen kuin ne tallennetaan hakemistoon, niin tarvitseeko tällainen hakemisto vähemmän muistitilaa kuin perinteisellä tavalla toteutettu hakemisto?
2) Jos hakija syöttää kyselyn hakusanat hakujärjestelmään perusmuodossa ja ne katkaistaan automaattisesti morfologisten tulkintaohjelmien avulla, niin ovatko automaattisesti katkaistuilla hakusanoilla saadun tulosjoukon saantiarvot, tarkkuusarvot tai molemmat keskimäärin korkeammat kuin perinteisellä tavalla hakijan itse katkaisemia hakusanoja käytettäessä?
3) Jos hakemiston sanat palautetaan perusmuotoon ja hakija antaa hakusanat perusmuodossa, niin ovatko näin saadun tulosjoukon saantiarvot, tarkkuusarvot tai molemmat keskimäärin korkeammat kuin perinteisellä tavalla hakijan itse katkaisemia hakusanoja käytettäessä?

Tutkimusongelma sekä siitä johdetut tutkimushypoteesit määritellään yksityiskohtaisemmin luvussa 6. Tutkimushypoteeseissa ja yleensäkin tässä tutkimuksessa käytetyt käsitteet määritellään luvussa 2.

1.4 FULLTEXT-projekti

Suomenkielisten tekstitietokantojen tallennus- ja hakutekniikat (FULLTEXT) -projektin tarkoituksena oli tuottaa perusselvitys siitä, miten suomen kielen morfologisten tulkintaohjelmien avulla voidaan lieventää tai ratkaista sellaisia tiedon tallennuksen ja haun ongelmia, jotka johtuvat suomen kielen erityispiirteistä. Tätä tarkoitusta varten rakennettiin erityinen testausympäristö, jossa samasta tekstiaineistosta tuotettiin joukko erilaisia hakemistoratkaisuja. Projekti oli siis luonteeltaan laboratorioympäristössä toteutettu evaluointitutkimus.

Hakujärjestelmien käytön helppoutta ei tässä tutkimuksessa varsinaisesti testattu eikä eri koejärjestelmiä annettu todellisten tiedontarvitsijoiden kokeiltaviksi, joskin asia pyrittiin ottamaan huomioon testejä laadittaessa ja analysoitaessa. Projektin lähtökohtana ei ollut rakentaa tuotantokäyttöön valmista hakujärjestelmää, vaan kokeellisin menetelmin vertailla eri vaihtoehtoja ja siten saada tietoa siitä, mikä tai mitkä morfologisten tulkintaohjelmien eri soveltamistavoista tarjoaisivat lupaavimmat jatkokehitysmahdollisuudet.

Eri tietokantoja rakennettaessa sovellettiin usealla eri tavalla suomen kielen morfogisia tulkintaohjelmia, ja näitä tietokantoja sekä niistä tehtyjen tiedonhakujen tuloksia vertailtiin toisiinsa. Testihaut pyrittiin toteuttamaan maksimaalisen tehokkaasti jokaisessa tietokannassa eli periaatteessa haettiin kaikki ne dokumentit, jotka kullakin kyselytyypillä olivat tuossa tutkimusympäristössä löydettävissä.

Vaikka FULLTEXT-projektissa tutkittiinkin perinteistä tiedonhakujärjestelmää, olivat projektissa käsitellyt ongelmat kuitenkin sellaisia periaatteellisen tason ongelmia, jotka koskevat muitakin suomenkielisiä, suuria tekstiaineistoja käsitteleviä tietojärjestelmiä.

Projektissa tehtiin seuraavat rajaukset: Testausympäristönä oli käänteishakemistoon ja Boolen logiikkaan perustuva, ns. perinteinen tiedonhakujärjestelmä. Tutkimusaineistona oli vapaamuotoinen, aihealueeltaan rajoittamaton teksti. Tekstillä ei siis ollut mitään rakennetta, jota sen käsittelyssä olisi voitu käyttää

hyväksi. Sen käsittelyssä ei myöskään voitu soveltaa mitään käsitemalleja tai erityissanastoja, kuten jonkin suppean aihealueen tekstiä käsiteltäessä.

Tulkintaohjelmien soveltamisen piti tekstin tallennusvaiheessa tapahtua täysin automaattisesti, koska suurten tekstimäärien vuorovaikutteiseen käsittelyyn ei yleensä ole käytännöllisiä tai taloudellisia mahdollisuuksia. Hakuvaiheessa tällainen rajaus ei kuitenkaan ole tarpeen, vaan silloin voidaan sallia, että hakujärjestelmä ongelmatilanteissa pyytää käyttäjältä tarkennuksia ja toimintaohjeita.

Suomen kielen tulkintaohjelmien vaikutusta hakunopeuteen ei projektissa tutkittu, koska ei ollut tarkoituksemukaista toteuttaa testiympäristöjä niin hyvin, että ne täyttäisivät tuotantokäytössä oleville hakujärjestelmille asetetut vaatimukset. Testaustarkoituksia varten riittäväksi katsottiin, että tulkintaohjelmat saatiin luotettavasti toimimaan hakujärjestelmissä ja että erilaisia hakemisto- ja kyselyratkaisuja voitiin kokeilla ja vertailla keskenään.

FULLTEXT-projektin testaukset toteutettiin kahdella eri hakujärjestelmällä: BASIS- ja APL-MINTTU-ohjelmistolla. Tässä opinnäytteessä käsitellään projektin BASIS-hakujärjestelmällä toteutettua osuutta. APL-MINTTU-testauksia on selostettu FULLTEXT-projektin loppuraportissa (Alkula & Honkela 1992).

2 KÄSITTEIDEN MÄÄRITTELY

2.1 Tiedonhaun käsitteet

Tiedon tallennus- ja hakujärjestelmä eli lyhyesti hakujärjestelmä (information storage and retrieval system) on tiedonhallintajärjestelmä, joka on suunniteltu erityisesti tekstimuotoisen tiedon tallennukseen ja hakuun. Hakujärjestelmää käytetään **kyselykielen** (command language, query language) avulla. Kyselykieleen kuuluu joukko **komentoja** (commands), jotka nimetään komentosanojen avulla ja joilla käyttäjä ohjaa tietokonetta suorittamaan halutun toiminnon. (Snow 1986; Järvelin 1995)

Tietokanta (database) on kokoelma tiettyä kohdetta (rajattua todellisuuden osa-aluetta) kuvaavia tietoja, jota ylläpidetään, täydennetään ja haetaan tietokonejärjestelmän avulla. Tietokanta koostuu joukosta **tietueita** (record). Tekstitietokannoissa tietueita voidaan nimittää myös **dokumenteiksi**. Sisällöllisesti tietueeseen on koottu tiettyä kohdetta (kuten henkilö, tuote tai tapahtuma) koskevien tietojen kokonaisuus yhdeksi käsiteltäväksi yksiköksi. Rakenteellisesti tietue koostuu joukosta nimettyjä kenttiä; **kenttä** (field) on tietovälineessä tai muistissa oleva rajallinen alue jonkin tiedon tallentamista varten. **Tiedosto** (file) puolestaan koostuu joukosta tietueita, jotka on järjestetty käsittelyä varten. Tiedosto on laaja-alaisempi käsite kuin tietokanta: tietokanta on tiedosto, mutta tiedosto ei aina ole tietokanta. (Atk-sanakirja 1990; Tietotekniikan sanasto 1990; Fidel 1987; Järvelin 1995.)

Tekstuaalisille tietokannoille on tyypillistä, että vapaamuotoista tekstiä sisältävät kentät ovat keskeisiä ja tiedonhaku perustuu useimmiten luonnollisen kielen sanojen käyttöön, vaikka osa tietueiden kentistä normaalisti sisältää myös määrämittaista ja -muotoista tietoa (kuten kirjoittajan nimen, päivämäärän, sisältö- tai kielikoodin jne.). Tällaisia tietokantoja ovat tyypillisesti kirjallisuus**viitetietokannat** (bibliographic database), esimerkiksi suomalainen KDOK-tietokantaperhe. Tämäntyyppisissä tietokannoissa julkaisun yksilöintitietojen (tekijä, julkaisun nimi/otsikko, julkaisija, ISBN-koodi jne.) lisäksi kuvataan sen tietosisältöä esimerkiksi tesauruksesta poimittujen kuvailutermien eli **indeksitermien**, tai luonnollisella kielellä kirjoitetun **tiivistelmän** avulla. **Kokotekstitietokannan**, lyhyesti **tekstikannan** (full-text database) tietueissa tekstillä on hallitseva osuus: dokumentin sisältöä kuvaa dokumentin teksti kokonaisuudessaan. Tällaisia tieto-

kantoja ovat lehtiartikkeleita, oikeuden päätöksiä tai lehdistötiedotteita tms. sisältävät tietokannat, esimerkiksi yhdysvaltalaiset WESTLAW ja LEXIS-NEXIS. (Tenopir & Ro 1990)

Kaupalliset tekstuaaliset tietokannat koostuvat yleensä **peräkkäistiedostosta** (sequential file, linear file, record file) ja **käänteistiedostosta** (inverted file) eli **hakemistosta**[1] (directory, index). Peräkkäistiedostossa ovat varsinaiset tietueet tietuenumeron mukaisessa järjestyksessä, yleensä tallennusjärjestyksessä. Hakemisto muodostetaan tekstiaineistosta yleensä siten, että kukin tekstiä sisältävässä kentässä esiintynyt sana - käytännössä kahden välilyöntimerkin tai muun määritellyn erottimen välissä esiintynyt merkkijono - poimitaan erikseen ja liitetään aakkosjärjestyksessä olevaan sanalistaan. Joissakin hakujärjestelmissä tuotetaan useampi kuin yksi hakemisto: esimerkiksi DIALOG-järjestelmässä muodostetaan aihesisältöön liittyvistä tekstikentistä (otsikko, tiivistelmä, indeksitermit) oma aakkosellinen hakemistonsa, ns. basic index, ja muista kentistä (kuten tekijä, lehden nimi yms.) muodostetaan omat, erilliset hakemistonsa. Toisissa hakujärjestelmissä (kuten BRS[2] ja WESTLAW) taas on vain yksi sekahakemisto, johon kaikkien (hakemistoon tallennettavien) kenttien sisällöistä poimitut merkkijonot on tallennettu. (Tenopir & Ro 1990)

Nykyisissä kaupallisissa hakujärjestelmissä, joita tässä tutkimuksessa nimitetään **perinteisiksi** hakujärjestelmiksi, sanat tallennetaan hakemistoon merkkijonoina, taivutusmuotoineen kaikkineen täsmälleen siinä muodossa kuin ne esiintyvät itse tietueessa. Tällaisesta merkkijonosta käytetään tässä ilmausta **hakemiston merkkijono** (index term). Luonnollisen kielen sanaa, jonka ilmentymiä (kuten eri taivutusmuotoja) tällaiset hakemiston merkkijonot ovat, kutsutaan **hakemistosanaksi** (Keskustalo 1994).

Kansainvälisissä tietokannoissa hakemistosta usein jätetään pois ns. **sulkusanat** (stop word), jotka ovat yleisiä, tiedonhaun kannalta merkityksettömiä sanoja. Eri hakujärjestelmien sulkusanamäärittelyt ovat erilaisia: DIALOG-

---

[1] Joissain määrittelyissä "hakemisto" voi tarkoittaa myös käänteistiedoston hakemistoa eli ns. sanakirjatiedostoa (dictionary file) (Järvelin 1995, s. 98); tässä tutkimuksessa hakemisto ja käänteistiedosto kuitenkin tarkoittavat samaa.

[2] Nykyään Ovid: vuonna 1994 CD Plus -niminen yritys osti BRS Onlinen ja muutti nimekseen seuraavana vuonna Ovid Technologies (URL: http://www.ovid.com/company/history.cfm)

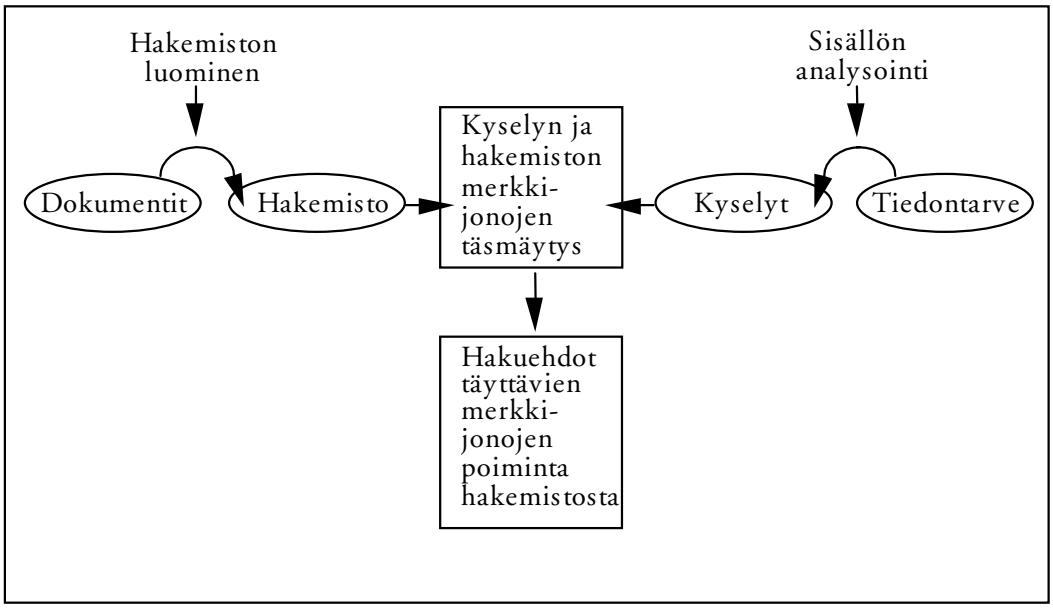
järjestelmässä sulkusanoja on yhdeksän (an, and, by, for, from, the, to, with, or[3]), BRS-järjestelmässä yli 70. (Tenopir & Ro 1990)

Tallennusvaiheessa hakemistosanoihin liitetään myös **osoite** tai osoitteet (address, posting) eli tieto siitä, missä tietueissa ne ovat esiintyneet. Yksinkertaisimmillaan osoite on tietueen järjestysnumero peräkkäistiedostossa, jolloin vain yleisesti tiedetään, että merkkijono esiintyy jossain kohtaa kyseistä dokumenttia. Kehittyneemmissä järjestelmissä osoite ilmaisee merkkijonon (hakemistosanan) tarkan sijainnin tekstikentässä. Tulostusvaiheessa järjestelmä hakee halutut tietueet esiin peräkkäistiedostosta hakemiston osoitetietojen perusteella.

Tiedonhaku, jatkossa myös lyhyemmin sanottuna **haku**, on prosessi, jonka tavoitteena on tiedontarpeiden tyydyttäminen. Tiedonhaussa pyritään löytämään tiedontarpeiden tyydyttämistä mahdollisimman hyvin palveleva dokumentti tai dokumenttijoukko. Hakuprosessia voidaan tarkastella monista eri näkökulmista. Suppeimmillaan hakuprosessia voidaan pitää pelkästään dokumenttien (käytännössä dokumenteissa esiintyneiden merkkijonojen) ja tiedontarvitsijan esittämästä hakupyynnöstä muokattujen formaalien esitysten **täsmäyttämisenä** (matching). Tällöin hakuohjelma käy läpi hakemistoa ja vertailee hakuavaimia hakemiston merkkijonoihin löytääkseen hakemistosta ilmaukset, jotka täsmäävät hakuavainten kanssa (kuva 1).

Tiedonhaku on kuitenkin monisyisempi prosessi kuin mekaaninen merkkijonojen täsmäyttäminen. Monet tiedonhakututkijat, esimerkiksi Belkin (1984), Schamber et al. (1990) ja Ingwersen (1992; 1996) ovatkin kritisoineet edelläkuvattua mekaanista mallia, joka unohtaa käyttäjän, ja esittäneet sen tilalle kognitiivista lähestymistapaa (cognitive viewpoint). Kognitiivisessa näkökulmassa huomio kiinnitetään haun yhteydessä esiintyviin erilaisiin ajattelu- ja tiedonkäsittelyprosesseihin: mitä tietämysrakenteita kullakin osanottajalla (dokumentin kirjoittaja, välittäjä, tiedontarvitsija) on, miten ne muuntuvat prosessissa, millaista tietoa vaihdetaan, mihin tiedontarve liittyy (konteksti) jne. Lisäksi otetaan huomioon, että ihmisillä ei välttämättä ole selkeää käsitystä tiedontarpeestaan eivätkä he näin ollen kykene kuvaamaan

[3] Tässä työssä noudatetaan merkintätapaa, jossa luonnollisen kielen ainekset (kuten sanojen esiintymät) on alleviivattu, esimerkiksi metsäteollisuuden. Kyselyissä käytetyt hakusanat puolestaan on kursivoitu: *metsäteollisuus*. Jos luonnollisen kielen ilmauksen jotain osaa on haluttu erityisesti korostaa, se on kursivoitu: *hevos*ta.



Kuva 1. Tekstin tallennus ja haku yksinkertaistettuna (mukaeltuna Salton 1989, s. 231).

tiedontarvettaan selkeästi ja johdonmukaisesti; elävässä elämässä hakupyynnöt ovat usein epämääräisiä ja kyselyt epätarkkoja.

Tiedonhankinnan (information seeking) tutkimuksessa selvitetään, mihin tietoa tarvitaan, miten sitä hankitaan ja miten käytetään (Järvelin 1995). Sukulaisuudestaan huolimatta tiedonhankinnan ja tiedonhaun tutkimukset ovat tähän asti olleet melko erillään toisistaan, vaikkakin monet tutkijat ovat olleet aktiivisia molemmilla alueilla (Ingwersen 1996; Vakkari 1999). Näiden kahden tutkimusalueen väliset yhteydet ovat kuitenkin lisääntymässä, koska on havaittu, että tiedonhankinnan prosessi vaikuttaa myös siihen, miten tietoa hakujärjestelmistä haetaan. Kun tiedontarvitsijan tietämys aihealueeseesta kasvaa, hänen laatimansa kyselyt muuttuvat, vaikka aihe pysyisikin samana; tämä johtuu siitä, että tiedontarvitsija kognitiivisen prosessin alussa tarvitsee erilaisia dokumentteja kuin prosessin lopussa. (Pennanen ja Vakkari 2000)

Tässä väitöskirjassa kuvattu tutkimus toteutettiin pelkästään tätä tutkimusta varten rakennetussa laboratorioympäristössä. Vaikka vaatimukselle tutkia vain todellisten tiedontarvitsijoiden aitoja kyselyjä osana aitoa tiedonhakuprosessia onkin pätevät perusteet, ei laboratorioympäristössä ole todellisia käyttäjiä. Erilaisten tietokanta- ja hakemistovaihtoehtojen toteuttaminen oikeassa, tuotantokäytössä olevassa tiedonhakujärjestelmässä olisi taloudellisesti mahdotonta. Ongelmana olisi myös saada riittävän vertailukelpoisia

tuloksia; tutkimusaineiston vakioiminen on vaikeaa toteuttaa muualla kuin laboratorioympäristössä. (Kristensen 1995)

Mekaanisen, järjestelmäsuuntautuneen tiedonhakututkimuksen kritiikki osuu siinä oikeaan, että hakujärjestelmien mekaanisilla vertailuilla ei tavoiteta tiedonhakuprosessin kaikkia ulottuvuuksia. Mutta toisaalta on todettava, että tiedonhaku on niin moniulotteinen prosessi, että sen tutkimiseen ja kuvaamiseen tarvitaan erilaisia menetelmiä (Pors 2000). Laboratoriotutkimus on yksi näistä menetelmistä; sen rajat on vain otettava huomioon tulosten tulkinnassa.

Tässä työssä ei käsitellä tiedonhaun dynaamista ja kognitiivista luonnetta tämän enempää eikä myöskään tiedonhakua prosessina. Aihetta on kuvattu perusteellisemmin muun muassa Järvelinin (1995, luku 2) oppikirjassa. Seuraavassa tiedonhakuprosessi kuvataan Järvelinin (1995) esittämän kolmitasoperiaatteen mukaisesti, jossa otetaan tiedonhaun kognitiivinen ja lingvistinen luonne huomioon. Tämä kuvaus on yleinen malli; yksittäisessä tiedonhakutilanteessa prosessi ei välttämättä etene täsmälleen siten kuin tässä kuvataan.

Tiedonhakutilanteessa tiedontarve eli käytännössä tiedontarvitsijan esittämä **hakupyyntö** aluksi analysoidaan **käsitetasolla** (conceptual level). Ensin suoritetaan käsiteanalyysi, jossa hakija tulkitsee hakupyynnön keskeiset käsitteet ja niiden väliset suhteet. Analysoinnin tuloksena on **käsitteellinen hakusuunnitelma**. Toisessa vaiheessa tämä käsitteellinen hakusuunnitelma käännetään **ilmaisutason hakusuunnitelmaksi**. Vastaavanlainen käsiteanalyysi tehdään myös silloin, kun dokumentteja kuvaillaan: indeksoija tulkitsee dokumentin keskeiset käsitteet ja näiden suhteet ja kääntää tämän käsiteanalyysin tuloksen indeksitermeiksi tai luonnollisen kielen sanoiksi. (Järvelin 1995)

Ilmaisutasolla (expression level) tarkastellaan käsitteiden ilmaisutapoja luonnollisella kielellä tai jollain erikoiskielellä (kuten dokumentaatiokielellä, jollaisia ovat muun muassa luokituskaavat, asiasanastot ja tesaurukset). Ilmaisutasolla käytetään hakupyynnössä esiintyneitä ilmauksia lähtökohtana, jonka perusteella kehitetään vaihtoehtoisia ilmaisuja - sanoja, sanontoja, indeksitermejä ja niin edelleen - hakupyynnön käsiterakenteen käsitteille. (Järvelin 1995) Jos esimerkiksi hakupyynnössä on käytetty luonnollisen kielen yhdyssanaa laktoosi-intoleranssi, sen rinnalle vaihto-

ehtoiseksi ilmaukseksi voidaan lisätä luonnollisen kielen yhdyssanojen sanaliitto maitosokerin imeytymishäiriö.

Hakupyynnön käsitteiden esityksiä ilmaisutasolla kutsutaan **hakuavaimiksi**. Hakusuunnitelma sisältää yhden tai useamman hakuavaimen. Luonnollisen kielen sanoihin perustuvia hakuavaimia voidaan sanoa myös **hakusanoiksi**. (Järvelin 1995) Tietueen tekstiä sisältäviin kenttiin (otsikko, tiivistelmä, tekstiosuudet) tai täsmällisemmin sanoen näissä kentissä esiintyviin sanoihin perustuvaa tiedonhakua kutsutaan **sanahauksi** (free-text retrieval).

Tietokoneisiin perustuva konkreettinen tiedonhaku tapahtuu aina **esiintymätasolla** (occurrence level). Tietokoneet eivät käsittele luonnollisen kielen sanoja sinänsä, vaan merkkijonoja. Esiintymätasolla ilmaisutason hakusuunnitelmasta muodostetaan tiedonhakujärjestelmän ymmärtämä formaali **kysely** (query). Käsitteellinen hakusuunnitelma kiinnittää huomion siihen, mitä tietoa haetaan, kysely puolestaan siihen, miten sitä haetaan. Esiintymätasolla ilmaisutason hakusuunnitelman hakuavaimet käännetään merkkijonoiksi. Kääntämisessä otetaan huomioon dokumenttien esitystavan, tietokantojen ja hakujärjestelmien tarjoamat mahdollisuudet ja rajoitukset. (Järvelin 1995)

Koska perinteiseen hakemistoon tallennetaan sanojen taivutusmuotoja, yksittäisestä sanasta esiintyy hakemistossa useita muotoja - jokainen eri taivutusmuoto on omanlaisensa merkkijono, **merkkijonovakio**. Jotta jokaista merkkijonovakiota eli hakuavaimen jokaista yksittäistä esiintymää hakemistossa ei tarvitsisi hakea erikseen, hakujärjestelmiin on kehitetty apuneuvoja, jolla useampi hakemiston merkkijono voidaan hakea yhdellä kertaa. **Merkkijonokaavio** on malli, joka täsmää useisiin merkkijonovakioihin, joilla on määrätyt yhteiset osat ja vaihtelua muissa osissa. Merkkijonon **katkaisu** (truncation) on merkkijonokaavioiden tavallisin erikoistapaus. (Järvelin 1995, s.199)

Käytännössä merkkijonon katkaisu toteutetaan **jokerimerkin** avulla. Jokerimerkki on symboli (esimerkiksi * tai ?), jonka tilalla saa esiintyä periaatteessa mikä tahansa merkki tai merkkijono. Kun jokerimerkki sijoitetaan hakusanan loppuun, hakujärjestelmä poimii hakemistosta kaikki ne merkkijonot, jotka alkavat samalla merkkiyhdistelmällä kuin hakusana jokerimerkkiin asti. Hakusanasta tällä tavoin muodostettua merkkijono-

kaaviota kutsutaan **katkaistuksi hakusanaksi**. Katkaistulla hakusanalla *auto** saataisiin muun muassa hakemiston merkkijonot auton, autossa, automaatti, ja autoliitto.

Kun hakija syöttää hakujärjestelmälle hakuavaimen, hakujärjestelmä täsmäyttää hakemiston merkkijonoja hakuavaimen kanssa. Jos hakemistosta löytyy yksi tai useampi hakuavaimen kanssa täsmäävä merkkijono, hakemiston merkkijonoon tai -jonoihin liitetyistä tietuenumeroista (osoitteista) muodostetaan numeroitu joukko, jota kutsutaan **tulosjoukoksi** (set, retrieval set) tai **hakutulokseksi**. Kyselyehdot täyttävää tietuetta sanotaan **osumaksi** (hit). Perinteisissä hakujärjestelmissä tulosjoukkoihin voidaan tiedonhaun kuluessa viitata tulosjoukon numerolla ja niitä voidaan kytkeä toisiin tulosjoukkoihin tai hakuavaimiin Boolen operaattoreiden ja läheisyysoperaattoreiden avulla. (Tenopir & Ro 1990, s. 63; Järvelin 1995, s. 101) Jos hakujärjestelmä ei löydä hakemistosta yhtään hakuavaimen kanssa täsmäävää merkkijonoa, eli ei saada yhtään osumaa, tiedonhaun tuloksena on **tyhjä (tulos)joukko**.

Boolen operaattoreita ovat JA-, TAI- ja EI-operaattorit. JA-operaattorilla (konjunktio) rajataan haun alaa; sillä ilmaistaan, että hakuavainten pitää esiintyä samassa dokumentissa. Eri hakuavaimilla saaduista tulosjoukoista siis tehdään leikkaus. Keskenään vaihtoehtoisia ilmauksia taas liitetään yhteen TAI-operaattorilla (disjunktio), ts. muodostetaan eri hakuavaimilla saatujen tulosjoukkojen yhdiste eli unioni. Riittää, että tietueessa on mainittu yksikin TAI-operaattorilla yhdistetyistä hakuavaimista. EI-operaattorilla (negaatio) määritellään sellaiset hakuavaimet, joita sisältäviä dokumentteja ei haluta mukaan, siis muodostetaan eri tulosjoukkojen erotus. (Järvelin 1995, luku 7.1)

Boolen operaattoreilla määritellyt operaatiot eivät aina tuota riittävän tarkkoja tuloksia. Pitkissä tekstidokumenteissa käy helposti niin, että JA-operaattorilla yhdistetyistä sanoista toinen esiintyy tekstin alussa ja toinen lopussa, eikä niillä käytännössä ole minkäänlaista keskinäistä yhteyttä. Jotta kyselyissä voitaisiin tarkemmin ilmaista, millaisessa tekstiyhteydessä hakusanan pitäisi esiintyä, monissa hakujärjestelmissä on käytössä **läheisyysoperaattorit**, joilla voidaan määritellä, miten lähellä toisiaan sanojen on sijaittava tekstissä. Läheisyysoperaattorilla voidaan esimerkiksi määritellä, että hakusanojen on esiinnyttävä samassa virkkeessä tai samassa kappaleessa. Jotta läheisyysoperaattoreita voitaisiin käyttää, on hakujärjestelmän

hakemisto jo tallennusvaiheessa laadittava siten, että osoitteet ilmaisevat sanojen tarkemman sijainnin dokumentin sisällä. (Tenopir & Ro 1990, s. 60 - 61)

Edellä on esitelty lyhyesti tiedonhaun yleiskäsitteitä; tarkempia määritelmiä ja tekstitiedonhaun erityisongelmien kuvauksia löytyy esimerkiksi tutkimuksen lähdekirjallisuudesta (Salton & McGill 1983; Salton 1989, Bain & al. 1989, Tenopir & Ro 1990; Sormunen & Alkula 1990; Glassco 1993; Järvelin 1995; Sormunen 1994 ja 2000).

2.2 Kielitieteen käsitteet

Morfologia on sanoja ja niiden rakennetta tutkiva kielitieteen haara. Sen tutkimuskohteena ovat morfeemit. **Morfeemi** on pienin kielellinen muoto, jolla on itsenäinen merkitys. **Vapaat** morfeemit voivat olla joko aina itsenäisiä (ja, mutta) tai sellaisia, jotka voivat esiintyä sekä itsenäisinä että yhdistyneenä toisiin morfeemeihin. Jälkimmäisiin kuuluvat ns. leksikaaliset morfeemit, esim. talo, ottaa, tuo. **Sidonnaiset** morfeemit ovat sellaisia, jotka eivät voi milloinkaan esiintyä yksin. Tällaisia ovat muun muassa **etuliitteet** eli prefiksit (*epä*kohta), **päätteet** eli suffiksit (talo*ssa*, kukka*nen*) ja **tunnukset** (esimerkiksi monikon i). (Hakulinen & Ojanen 1976) Synteettiselle kielelle, kuten suomelle, on ominaista, että vapaaseen morfeemiin (useimmiten sen loppuun) liittyy erilaisia sidonnaisia morfeemeja. Analyyttisessä kielessä, jollainen esimerkiksi englanti on, morfeemit ovat enimmäkseen itsenäisiä.

Perinteisesti sana on määritelty pienimmäksi merkitystä kantavaksi yksiköksi, johon lause voidaan jakaa (Hakulinen & Ojanen 1976). Tässä tutkimuksessa **sana**-käsite on lähinnä **lekseemin** synonyymi eli abstraktio, joka kattaa sanan kaikki eri esiintymät. **Sananmuodot** eli **saneet** ovat puheessa ja kirjoituksessa esiintyviä (reaalistuvia) sanan (lekseemin) esiintymiä. Sanan **paradigman** muodostavat **perusmuoto** ja **taivutusmuodot**. Suomen kielessä nominien perusmuodoksi määritellään yksikön nominatiivi (talo, se, kultainen) ja verbien perusmuodoksi I infinitiivi (juosta, uneksia) (Karlsson 1994, s. 171)

Sanan **taivutus** on morfologisten keinojen käyttöä sananmuotojen, taivutusmuotojen muodostamiseksi lekseemistä. Taivutusmuodot ilmaisevat kieliopillisia suhteita, taivutetun sanan suhdetta lauseen muihin sanoihin (Karlsson 1994). Perusmuodolla ja taivutusmuodoilla on yhteinen alkuosa, jota

nimitetään **vartaloksi**. Vartalo ei välttämättä ole kaikissa paradigman muodoissa täsmälleen samanlainen, vaan siinä voi esiintyä äännevaihteluita (esimerkiksi *hevose*na, *hevos*ta).

Suomenkielisen sanan vartaloon voi liittyä seuraavanlaisia suffikseja (Leino 1991):

1. **Pääte** ilmaisee sanan suhteen lauseen muihin sanoihin. Sijapäätteet (kuten -ssa, -sta, -na) liittyvät nominin ja persoonapäätteet (kuten -n, -mme, -tte, -vat) verbin vartaloon. Muita varsinaisia päätteitä ei suomessa ole.
2. **Johdin** on suffiksi, jonka avulla muodostetaan uusi sana, kuten: kahvi + la -> kahvila, laiva + sto -> laivasto
3. **Tunnuksen** avulla muodostetaan vartalosta samaan paradigmaan uusi alivartalo. Esimerkiksi nominivartaloon voi liittyä monikon tunnus ja verbivartaloon aikamuodon tunnus.
4. **Liitteet** ovat omistusliitteitä (kuten -ni, -nsa, -mme, -nne) tai liitepartikkeleita (kuten -ko, -pa, -kin, -han)

Perussana on yhdistämätön vapaa morfeemi (esimerkiksi kala, keitto). **Johdos** on toisesta sanasta johtimen avulla muodostettu sana. Johtaminen voi toisinaan muuttaa sanan sanaluokkaa, mitä ei koskaan tapahdu taivutuksessa. Esimerkiksi -minen-johtimella verbistä johdetaan substantiivi: tule+*minen* (Karlsson 1982). Sanan **kanta** on se johdetun sanan osa, joka jää jäljelle, kun johdin tai johtimet on erotettu (*kala*+sta+ja). **Yhdyssana** on kahdesta tai useammasta perussanasta tai johdoksesta muodostettu sana (esimerkiksi kalakeitto).

Homonymia on ilmiö, jossa eri asiaa merkitsevät sanat ovat kirjoitus- ja äänneasultaan identtiset (Hakulinen & Ojanen 1976). Homonymian alalaji on homografia, jossa vain sanojen kirjoitusasu on sama. Ääntämisestä riippuen esimerkiksi sananmuoto hauissa voi olla joko hauki- tai haku-sanan taivutusmuoto. **Täydellisessä homonymiassa** kahden eri sanan paradigmat ovat täysin yhtenevät, eli ne ovat kaikissa muodoissaan keskenään identtisiä, kuten vaara 'riski, uhka' ja vaara 'kukkula, mäki, vuori'. **Sananmuotohomonymiassa** vain osa sananmuodoista on identtisiä, esimerkiksi satoja voi olla joko sata- tai sato-sanan esiintymä (Penttilä 1975; Laalo 1990, s. 18). Hjorth (1987b) käyttää näistä nimitystä **osahomografi**.

Homonymialle lähekkäinen ilmiö on **polysemia** eli **monimerkityksisyys**, jossa samalla sanalla on kaksi tai useampia lähekkäisiä merkitysvariantteja, esimerkiksi (ihmisen, kirjan, järven, yön jne.) selkä tai (ihmisen, tien) poski (Hakulinen & Ojanen 1976; Karlsson 1994, s. 197). Homonymian ja polysemian erona on, että homonymia on kahden tai useamman sanan välinen suhde, kun taas polysemia on yksittäisen sanan ominaisuus. Tosin näiden kahden ilmiön erottaminen on usein käytännössä vaikeaa, vaikka ero teoreettisesti olisikin selvä (Penttilä 1975). Sitä paitsi kielen kehittyessä tapahtuu siirtymistä polysemiasta homonymiaan, kun sanan merkitysvivahteiden välinen ero ajan myötä kasvaa riittävästi (Laalo 1989).

Luonnollisessa kielessä homonymia ja erityisesti polysemia ovat hyödyllisiä ilmiöitä: on taloudellista, että jokaiselle käsitteelle ja merkitysvivahteelle ei tarvitse olla omaa ilmausta, vaan samaa ilmausta voidaan joustavasti käyttää eri tarkoitteisiin. Kuulija tai lukija osaa asiayhteyden eli **kontekstin** perusteella päätellä ilmauksen oikean tulkinnan. Ongelmana onkin, että suomen kielen morfologiset tulkintaohjelmat käsittelevät vain erillisiä sananmuotoja ilman kontekstia, jolloin ne eivät kykene päättelemään, mikä vaihtoehtoisista tulkinnoista on oikea.

Sananmuotoja, joista on mahdollista tehdä useampi kuin yksi tulkinta, sanotaan **monitulkintaisiksi**. Sananmuotohomonymian määrä vaihtelee kielittäin. Suomenkielisessä tekstissä noin 15 % sananmuodoista on monitulkintaisia, kun taas englannissa noin puolet kirjoitetun tekstin sananmuodoista on ainakin kaksiselitteisiä homografeja ja ruotsinkielisessä tekstissä puolestaan yli 65 % (Karlsson 1994, s. 80). Oikean tulkinnan etsimistä (esimerkiksi eri vaihtoehtoja yksi kerrallaan pois karsien) sanotaan **disambiguoinniksi** eli yksiselitteistämiseksi.

Homonymian vastakohta on **synonymia**, jossa kahdella äänne- ja kirjoitusasultaan erilaisella ilmauksella on sama tarkoite ja merkitys. Kahden tai useamman sanan sanotaan olevan synonyymeja, jos lauseilla, joissa ne on korvattu toisillaan, on sama merkitys (Hakulinen & Ojanen 1976).

2.3 Tutkimuksen erityiskäsitteet

Suomen kielen morfologisten tulkintaohjelmien hyödyntäminen tiedonhaussa on alue, jolle kaikilta osin ei löydy valmiita käsitteitä sen enempää kielitieteen kuin tiedonhakututkimuksenkaan alueelta. Siksi tässä tutkimuksessa on tarpeen määritellä joukko erityiskäsitteitä:

Perusmuoto-ohjelma on morfologinen tulkintaohjelma, joka pystyy tuottamaan sille syötetyistä sananmuodoista näiden perusmuodot. Suomen kielen perusmuoto-ohjelmia ovat Morfo ja Twol (ks. luku 5.2.2).

Taivutusvartalo-ohjelmat tai lyhyemmin **vartalo-ohjelmat** puolestaan tuottavat niille syötetyistä perusmuodoista taivutusvartalot. Suomen kielen taivutusvartaloita tuottavat Finstems ja Hahmotin (luku 5.2.1). Nämä ohjelmat muodostavat sanan kaikki erilaiset mahdolliset vartalot, jotka yhdessä kattavat sanan paradigman eri muodot. Tässä määritellään, että taivutusvartaloissa voi olla jäljellä osia sananmuotojen päätteistä tai tunnuksista (kuten osake-, osakkee-, osakkei-), kun taas vartaloon ei tällaisia pääteaineksia kuulu.

Perusmuotohakemisto on tiedonhakujärjestelmän hakemisto eli käänteistiedosto, joka on muodostettu perusmuoto-ohjelman avulla siten, että tallennettavan tekstin sananmuodot on ennen hakemistoon tallennusta palautettu perusmuotoon (eli **perusmuotoistettu**). Sananmuodot, joita perusmuoto-ohjelma ei ole pystynyt tulkitsemaan, ovat **tunnistamattomia** sananmuotoja ja tallennetaan omaan hakemistoonsa, jota nimitetään **tunnistamattomien sananmuotojen hakemistoksi**. Tunnistamattomia sananmuotoja ei käytännössä tarvitse tallentaa fyysisesti erilliseen hakemistoon, vaan ne voivat olla samassa hakemistossa kuin perusmuodotkin. Tällaisessa vaihtoehdossa kuitenkin on olennaista, että tunnistamattomat sananmuodot pystytään tarvittaessa erottamaan hakemistossa perusmuoto-ohjelman tunnistamista sanoista. Tämä voidaan tehdä mahdolliseksi esimerkiksi merkitsemällä tunnistamatta jääneet sananmuodot jollain symbolilla.

Toisaalta varsinainen perusmuotohakemistokin voi vahingossa sisältää myös sanojen taivutusmuotoja: näin voi käydä silloin, kun perusmuoto-ohjelma tulkitsee jonkin sananmuodon väärin tai tuottaa ylimääräisiä tulkintoja (esimerkiksi perusmuotohakemistosta löytyvä kokkolasta-sananmuoto viittaisi siihen, että Kokkola-nimen taivutusmuoto on tulkittu yhdyssanaksi, joka koostuu kokko- ja lasta-perusmuodoista).

Milloin tässä tutkimuksessa ei erikseen täsmennetä, tarkoittaa "perusmuotohakemisto" sitä hakemistokokonaisuutta, joka muodostuu sekä varsinaisesta perusmuotohakemistosta että sitä täydentävästä tunnistamattomien sanojen hakemistosta.

Suomen kielen perusmuoto-ohjelmat pystyvät haluttaessa perusmuotoistamisen lisäksi jakamaan yhdyssanat osiin. **Ositettu perusmuotohakemisto** on hakemisto, joka sisältää perusmuotojen lisäksi myös yhdyssanan osat. (Yhdyssanojen osien tallennus voidaan tehdä monilla eri tavoilla; ks. luku 7.2.)

Taivutusmuotohakemisto sisältää tekstissä esiintyneet sananmuodot taivutusmuotoisina. Perinteisen hakujärjestelmän hakemisto on tälläinen hakemisto. Itse asiassa myös tunnistamattomien sananmuotojen hakemistoa pitäisi sanoa taivutusmuotohakemistoksi, koska sekin sisältää sananmuodot taivutusmuotoisina. Selkeyden vuoksi tässä tutkimuksessa kuitenkin rajataan termi "taivutusmuotohakemisto" tarkoittamaan yksinomaan perinteisen hakujärjestelmän koko hakemistoa, kun taas sen osajoukkoa, (ositettua) perusmuotohakemistoa täydentävää hakemistoa sanotaan "tunnistamattomien sananmuotojen hakemistoksi".

Hakijalla tarkoitetaan tässä tutkimuksessa yleensä tiedonhaun ammattilaista, joka osaa käyttää hakujärjestelmiä ammattitaitoisesti. **Satunnainen hakija** puolestaan on henkilö, joka ei tee tiedonhakuja ammatikseen, vaan ainoastaan silloin tällöin käyttää hakujärjestelmää.

Hakijan katkaisema hakusana on sellainen merkkijonokaavio, jolla pyritään löytämään kaikki sanan taivutusmuodot, mutta samalla mahdollisuuksien mukaan myös minimoimaan väärät osumat eli sellaiset hakemiston merkkijonot, jotka vain sattumalta täsmäävät katkaistun hakusanan kanssa. Kaikilla sanoilla optimaalista tulosta ei välttämättä saavuteta vain yhdellä merkkijonokaaviolla, mikäli eri taivutusmuotojen vartalot poikkeavat toisistaan huomattavasti (esimerkiksi yö -öitä).

Taivutusvartalo-ohjelmien katkaisema hakusana on **taivutusvartalo**. Taivutusvartaloita voi sanalla olla yksi tai useampi, ja sillä (niillä) pyritään kattamaan sanan kaikki erilaiset esiintymät. Pyrkimyksenä on tuottaa mahdollisimman pitkiä - ja kielitieteellisesti virheettömiä - vartaloita, sillä mitä pidempi merkkijonokaavio on, sitä harvempi hakemiston merkkijono täsmää sen kanssa. Kun merkkijonokaavion pituus kasvaa esimerkiksi yhdellä merkillä, kaikki ne hakemiston merkkijonot, joissa tuon merkin sijalla on jokin muu merkki, karsiutuvat pois. (Taivutusvartaloa on tämän aihepiirin aiemmissa suomalaisissa tutkimuksissa kutsuttu myös hakuvartaloksi, siis tiedonhaussa käytetyksi vartaloksi; esim. Nurminen 1986, Hjorth 1987b.)

Tiedonhaussa TAI-operaattorilla yhdistettävät vaihtoehtoiset hakusanat eivät välttämättä ole todellisia synonyymejä, vaikka niitä haussa kohdellaan synonyymisinä ilmauksina. Yksi tällainen kvasisynonyymien tai "hakunyymien" luokka ovat kantasanan ja sen johdosten muodostama joukko, esimerkiksi kirja, kirjoittaa, kirjoitus, kirjoittaminen, kirjoittaja, kirjasto jne. Tässä tutkimuksessa tällaisesta hakusanan kantasanan ja johdosten muodostamasta joukosta käytetään nimitystä **johdosperhe**.

Vakiokysely on kysely, jonka suorittamisessa suomen kielen morfologisten tulkintaohjelmien soveltaminen ei tuota erityisongelmia. **Ongelmakysely** on tulkintaohjelmien kannalta hankala kysely esimerkiksi siksi, että hakusanat ovat siinä taivutusmuotoisia tai puuttuvat perusmuoto-ohjelman sanakirjasta.

3 TIEDONHAUN TULOKSELLISUUDEN MITTAAMINEN

3.1 Relevanssin määrittely

Tiedonhaussa pyrkimyksenä on löytää kaikki tai riittäväksi katsottu määrä dokumentteja, jotka tyydyttävät tiedontarpeen (hyvä saanti), ja samalla pitää ei-toivottujen dokumenttien määrä mahdollisimman pienenä (hyvä tarkkuus). Se on vuorovaikutteinen prosessi, jossa kyselyä muokkaamalla pyritään siihen, että relevanttien dokumenttien osuus tulosjoukossa lisääntyy - joko lisäämällä relevanttien dokumenttien määrää tai poistamalla epärelevantteja dokumentteja.

Suppeasti määriteltynä **relevanssi** ilmaisee, miten hyvin löydetty tietue vastaa hakupyynnössä esitettyä tiedontarvetta. Relevanssi on kuitenkin hyvin kiistanalainen käsite. Vaikka sillä on ollut olennainen merkitys tiedonhakujärjestelmien kehittämisessä ja arvioinnissa, relevanssin käsitettä ei vielä tähän päivään mennessä ole saatu täsmällisesti määriteltyä, olkoonkin että informaatiotutkijat ovat kiistelleet asiasta 1950-luvun lopulta asti. Välillä relevanssiin liittyvät kysymykset ovat jääneet alan keskustelussa sivummalle, mutta 1980-luvun lopulla ja 1990-luvun alussa ne nousivat jälleen keskeiseksi keskusteluaiheeksi. (Schamber et al. 1990; Froehlich 1994)

Yksi ongelmista on, miten erotetaan - tai että pitäisikö ylipäätään erottaa - toisistaan "objektiivinen" aiheenmukaisuus (topicality, aboutness, subject relatedness) ja "subjektiivinen" käyttökelpoisuus käyttäjän kannalta (user relevance, pertinence, utility). Esimerkiksi Swanson (1977) totesi, että varhaisissa tiedonhakututkimuksissa dokumentit on saatettu arvioida relevanteiksi tai epärelevanteiksi jostain muustakin syystä kuin niiden aiheenmukaisuuden perusteella. Näiden tutkimusten arviointiperusteista ei kuitenkaan ole mitään tietoa, koska relevanssin käsitettä ei varsinaisesti ollut määritelty eikä relevanssiarvioitsijoiden henkilökohtaisiin ominaisuuksiin alan varhaisissa tutkimuksissa erityisesti kiinnitetty huomiota. Oli esimerkiksi mahdollista, että dokumentti käsitteli haluttua aihetta, mutta arvioija katsoi sen epärelevantiksi siksi, ettei se sisältänyt mitään arvioijalle uutta tietoa.

Harter (1992) puolestaan esitti joukon esimerkkejä, joissa dokumentti voi olla tiedontarvitsijan kannalta hyvinkin relevantti, vaikka ei lainkaan käsittele hakupyynnössä ja kyselyssä määriteltyä aihepiiriä. Jos katsotaan, että

tiedontarve on kognitiivinen tila, jossa etsitään tietoa omien ajatusmallien rakentamiseksi ja uusien ideoiden ja yhteyksien luomiseksi, saattavat nimenomaan määrätyn aihealueen ulkopuoliset (joskin sitä jollain tavalla sivuavat) dokumentit olla tiedontarpeen kannalta relevanteimpia, koska ne todennäköisemmin tarjoavat tiedontarvitsijalle uutta tietoa ja uusia ajatusmalleja kuin sellaiset aihealuetta varsinaisesti käsittelevät dokumentit, joiden asiasisältö voi olla tiedontarvitsijalle jokseenkin tuttu. Myös Spink et al. (1998) huomauttavat, että vaikka tiedontarvitsijat haluavatkin relevantteja dokumentteja, ovat myös osittain relevantit dokumentit heille käytännössä tärkeitä, koska niissä voi olla uusia näkökulmia.

Regazzin (1988) tutkimuksen perusteella voidaan pohtia, onko aiheenmukaisuuden ja hyödyllisyyden välinen ero lähinnä relevanssitutkijoiden käsitteenmäärittelyyn liittyvä ongelma. Kun Regazzi antoi samat dokumentit arvioitavaksi kahdelle eri koehenkilöryhmälle ja pyysi toisen ryhmän arvioimaan dokumenttien aiheenmukaisuutta ja toisen ryhmän dokumenttien hyödyllisyyttä, ryhmien antamat relevanssiarviot eivät juuri poikenneet toisistaan. Sen sijaan itse koehenkilöiden ominaisuuksilla (kuten aihealueen asiantuntemuksella) oli suuri vaikutus siihen, mikä dokumentti katsottiin relevantiksi ja mikä ei.

Laajassa katsausartikkelissaan Schamber et al. (1990) toteavat perinteisen näkemyksen relevanssista olevan liian suppean. Tiedontarvitsija ja tiedonhakujärjestelmä on katsottu erillisiksi ilmiöiksi ja tiedontarvitsijan tehtävänä on ollut vain passiivisesti reagoida hakujärjestelmän tuottamiin hakutuloksiin eli lajitella tulosjoukon tietueet relevantteihin ja epärelevantteihin. Vaikka Schamber et al. eivät halua varsinaisesti kieltää tällaisen suoraviivaisen mallin (source-to-destination) käyttökelpoisuutta tiedonhakujärjestelmien kehittämisessä ja arvioimisessa, he toteavat sen liian rajoittuneeksi itse relevanssikäsitteen kannalta. Järjestelmäsuuntautuneesta tutkimuksesta pitäisi siirtyä käyttäjäkeskeiseen tutkimukseen, jossa tiedontarvitsijoiden tilanne ja henkilökohtaiset ominaisuudet otetaan huomioon. Dokumentin relevanssi ei ole staattinen, vaan dynaaminen ominaisuus, joka muuttuu koko ajan sen mukaan kuin tiedontarvitsija omaksuu uutta tietoa. Esimerkiksi jonkin tutkimusprojektin alkuvaiheessa relevantiksi katsottu dokumentti ei aina ole sitä tutkimuksen loppuvaiheessa. Relevanssiarvio ei siis ole mikään lopullinen totuus, vaan tietyllä hetkellä tehty tilannekatsaus.

Schamber et al. (1990) peräänkuuluttivatkin tutkimusta, joka selvittäisi niitä perusteita, joiden perusteella tiedontarvitsijat arvioivat dokumentteja. Näi-

den perusteiden avulla voitaisiin määritellä uusi, moniulotteinen relevanssikäsite. Froehlich (1994) kuitenkin arvelee, että relevanssille tuskin löytyy yhtä yleisesti hyväksyttyä määritelmää - jos se olisi mahdollista, sellainen olisi kuluneiden vuosikymmenten aikana jo pitänyt saada aikaiseksi. Toisaalta on mahdollista löytää ja systematisoida niitä eri perusteita, joilla tiedontarvitsijat dokumentteja arvioivat. Tällaisia ovat esimerkiksi julkaisun tai kirjoittajan arvovaltaisuus sekä dokumentin tuoreus ja saavutettavuus. Froehlich toteaa myös, että vaikka aiheenmukaisuus yksinään ei ole riittävä, se on kuitenkin välttämätön peruste, joka muodostaa relevanssi(arvio)n ytimen. Niinpä aiheenmukaista tiedonhakua sekä voi että pitääkin kehittää. Tiedonhakujärjestelmiä tulisi kehittää sellaisiksi, että ne voisivat automaattisesti suodattaa tai asettaa paremmuusjärjestykseen dokumentit sen perusteella, mitä ominaisuuksia tiedontarvitsijat ovat määritelleet tärkeiksi; nämä perusteet siis olisivat samoja, joita tiedontarvitsijat itse soveltavat dokumenttien relevanttiutta arvioidessaan.

Schamberin ja kumppaneiden (1990) sekä Froehlichin (1994) kaipaamia perusteita ovat myöhemmin listanneet muun muassa Cosijn & Ingwersen (2000). He luokittelivat relevanssin ilmenemismuodot viiteen ryhmään:

1. laskennallinen relevanssi (system/algorithmic): kuvaa kyselyn ja kohteiden (information objects; esimerkiksi tekstien) välistä suhdetta; löytyvätkö niistä samat elementit (kuten tietty merkkijono)
2. aiheenmukainen (topical): kuvaa kyselyssä esitetyn aiheen ja kohteissa esitetyn aiheen suhdetta; käsittelevätkö ne samaa aihetta
3. kognitiivinen (cognitive): kuvaa tiedontarvitsijan tietämyksen ja tiedontarpeen suhdetta kohteisiin; ymmärtääkö tiedontarvitsija kohteen relevantiksi
4. tilanteenmukainen (situational): kuvaa kohteen soveltuvuutta tehtävänratkaisuun, päätöksenteon tukena tms.; auttaako kohde tiedontarvitsijaa ratkaisemaan tietyn tehtävän
5. sosio-kognitiivinen (socio-cognitive): kuvaa tiedontarvitsijan tai yhteisön kykyä hyödyntää kohdetta; suhde tiedontarvitsijan kokemukseen, yleisiin alan käytäntöihin ja tieteellisiin paradigmoihin

Yllälistatuista relevanssityypeistä vain ensimmäinen ei ole kontekstisidonnainen: loput neljä ovat subjektiivisia eli riippuvia tiedontarvitsijan ominaisuuksista, tilanteesta yms. Kuvauksen perusteella voidaan siis todeta, että

FULLTEXT-tutkimus ei ole käyttäjäkeskeinen, vaan lukeutuu perinteiseen järjestelmäsuuntautuneeseen tutkimukseen, jossa relevanssi on käsitetty suppeassa, Cosijnin ja Ingwersenin jaottelun mukaan algoritmisessa mielessä.

Relevanssin käsitteen syvällisempi määrittely ja analyysi rajataan tämän tutkimuksen ulkopuolelle lähinnä resurssisyistä ja koska se ei varsinaisesti sisältynyt tämän tutkimuksen ongelmanasetteluun. Toisaalta anglosaksisen sanonnan mukaan voi sanoa, että "kotityöt on tehtävä" ennen kuin päästään niin pitkälle, että hakujärjestelmää voidaan kehittää todellisten käyttäjien kanssa.

Robertsonin ja Hancock-Beaulieun (1992) mukaan laboratoriotestit ja operationaaliset testit yhdistetään parhaiten siten, että tutkimuksen alussa suuri joukko muuttujia testataan laboratoriossa ja tämän jälkeen siirrytään operationaalisiin testeihin, joissa käytetään pienempää joukkoa muuttujia. FULLTEXT-projektin lähtökohtaa tuo periaate kuvaakin hyvin. Erilaisten vaihtoehtoisten mallien testauksella haluttiin selvittää, onko morfologisten analyysiohjelmien soveltaminen edes teknisellä tasolla mielekästä ja hyödyllistä - jos jokin tutkimusympäristö osoittautuu jo tekniseltä kannalta toteutuskelvottomaksi, sen tutkiminen voidaan lopettaa ja myöhempi tutkimus kohdentaa lupaavimmiksi todettuihin vaihtoehtoihin.

3.2 Saannin ja tarkkuuden laskeminen

Tiedonhaun tuloksellisuutta on perinteisesti mitattu saanti- ja tarkkuusarvoilla, jotka perustuvat dokumenttien relevanssin määritykseen. Saantia (recall) ja tarkkuutta (precision) on pidetty tiedonhaun laadun mittareina.

Näiden laatumittarien lisäksi voidaan mitata myös määrällisiä ominaisuuksia kuten tulosjoukon kokoa. Tulosjoukon koko ei nimittäin saisi ylittää hakijan asettamaa raja-arvoa eli selauspistettä. Blairin (1990) mukaan hakija ei tutki tulosjoukon sisältöä ennen kuin tulosjoukko on saatu hakijan mielestä kohtuullisen kokoiseksi. Muuten hakija lisää rajoittavia hakuavaimia kyselyyn, kunnes saa supistettua dokumenttien määrän alle selauspisteen.

Blair (1990) kuitenkin toteaa ongelmaksi sen, että hakija ei välttämättä osaa supistaa tulosjoukkoa parhain mahdollisin keinoin, vaan lisää vääränlaisia rajoitteita ja sen seurauksena epähuomiossa karsii pois myös käyttökelpoisia dokumentteja. Toisaalta kyselyyn ei kannata lisätä sellaisia vaihtoehtoisia (TAI-operaattorilla kytkettyjä) hakusanoja, jotka tuottavat tulosjoukkoon

vain muutaman tai ei yhtään uutta relevanttia dokumenttia, mutta samalla kasvattavat tulosjoukon koon moninkertaiseksi.

Ideaalitapauksessa kyselyä muokkaamalla saadaan tulosjoukossa sekä saanti- että tarkkuusarvot nousemaan; käytännössä ne tavallisesti ovat toisilleen käänteiset: saannin parantuessa tarkkuus huononee ja päinvastoin.

Määrittelyn mukaan **saanti** *(R_i)* ilmaisee, miten suuri osuus kaikista tietokannassa olevista kyselyn *i* kannalta relevanteista dokumenteista (t_i) saatiin mukaan kyselyn *i* tulosjoukkoon. Saanti lasketaan kaavan (1) mukaisesti, jossa (r_i) on relevanttien dokumenttien määrä *i*:n tulosjoukossa:

$$(1) \qquad R_i = \frac{r_i}{t_i}$$

Tarkkuus *(P_i)* taas ilmaisee, kuinka suuri osuus tulosjoukon dokumenteista on relevantteja *(r_i)* suhteessa koko tulosjoukkoon eli kaikkiin kyselyn *i* tuloksena saatuihin dokumentteihin *(n_i)*. Mitä vähemmän epärelevantteja dokumentteja, sen parempi. Tarkkuus lasketaan kaavan (2) mukaan:

$$(2) \qquad P_i = \frac{r_i}{n_i}$$

Suurissa tutkimustietokannoissa saantiarvot yleensä lasketaan ns. **suhteellisen** eikä **absoluuttisen** saannin periaatteella. Absoluuttinen saanti ilmaisee, miten suuri osuus relevantteja dokumentteja on onnistuttu löytämään suhteessa kaikkiin tietokannassa oleviin relevantteihin dokumentteihin. Absoluuttisen saannin laskeminen edellyttäisi, että tietokannan jokaisen artikkelin relevanssi kunkin kyselyn suhteen olisi määritelty. Suurissa tutkimustietokannoissa tämä kuitenkin on käytännössä mahdotonta.

Suhteellisen saannin laskemiseen riittää, että tietokannasta määritellään jokin rajallinen ja riittävän kattava osajoukko, joka sitten edustaa tietokannan kaikkia relevantteja artikkeleita ja johon kunkin kyselyn tuloksena saatua tulosjoukkoa suhteutetaan. FULLTEXT-projektissa tällainen saantikanta muodostettiin niputtamalla kaikista vertailtavista hakemistovaihtoehdoista kaikkia eri kyselytyyppejä käyttäen tietokannasta löydetyt relevantit artikkelit yhteen. Kunkin yksittäisen kyselytyypin saantiarvo mitattiin laskemalla, kuinka monta saantikannan artikkelia tällä kyselytyypillä onnistuttiin löytämään. Jos jollain kyselytyypillä saadaan tulokseksi kaikki saantikannan artikkelit, se saa saantiarvokseen 100 %.

Yksittäisten kyselyjen tuloksina saatujen tulosjoukkojen saanti- ja tarkkuusarvojen perusteella tutkimuksissa yleensä lasketaan, miten tietyntyyppisten kyselyjen hakutulokset suhteutuvat vertailtavien kyselytyyppien vastaaviin tuloksiin. Tämä vertailu voidaan tehdä joko ns. mikro- tai makrokeskiarvon perusteella. (Tague-Sutcliffe 1992)

FULLTEXT-projektissa kunkin kyselytyypin keskimääräinen tarkkuus ja suhteellinen saanti laskettiin makrokeskiarvon periaatteella. Tällöin jokaisen haun tarkkuus- ja saantiarvot lasketaan erikseen ja näistä arvoista otetaan edelleen keskiarvo. Näin jokainen kysely saa yhtäläisen painoarvon kyselytyypin tehokkuuden laskemisessa. Tämä laskutapa sopii vertailuihin, joissa eri kyselyjen tuloksina saadut tulosjoukot voivat olla hyvin erikokoisia, kuten tässä tutkimuksessa oli laita. (Tague-Sutcliffe 1992)

Makrokeskiarvo (suhteellisesta) saannista lasketaan kaavalla:

$$(3) \qquad R_{makro} = \frac{\sum_{i=1}^{q} \frac{r_i}{t_i}}{q}$$

missä r_i on kyselyn i tulosjoukossa olevien relevanttien dokumenttien määrä, t_i kaikkien relevanttien dokumenttien määrä (saantikanta) kyselyssä i ja q kyselyjen kokonaismäärä.

Makrokeskiarvo tarkkuudesta lasketaan kaavalla:

$$(4) \qquad P_{makro} = \frac{\sum_{i=1}^{q} \frac{r_i}{n_i}}{q}$$

missä n_i on kyselyn i tulosjoukon koko eli dokumenttien määrä tulosjoukossa (muut symbolit kuten edellä).

4 KIELITIETEEN HYÖDYNTÄMINEN TIEDONHAKUTUTKIMUKSESSA

4.1 Kielen osajärjestelmät

Luonnollinen kieli on monitasoinen (monikerroksinen) järjestelmä. Se jakautuu osajärjestelmiin, jotka ovat keskenään monenlaisissa suhteissa. Osajärjestelmiin jaottelu vaihtelee eri kirjoittajilla ja eri ajankohtina. Esimerkiksi Karlsson mainitsee aiemmassa oppikirjassaan (1976, s. 44) neljä tasoa: fonologian, morfologian, syntaksin ja semantiikan. Näistä kolme ensimmäistä ovat kieliopin muodolliset osajärjestelmät, kun taas semantiikka liittyy sisältöön (merkitysrakenne). Doszkocs (1986) puolestaan esittelee kuusi eri tasoa: fonologian, morfologian, leksikon, syntaksin, semantiikan ja pragmatiikan. Karlssonillakin (1976) pragmatiikka on mainittu lausesemantiikkaa käsittelevässä luvussa. Smeatonin (1991) luettelosta löytyvät fonologia, leksikko, syntaksi, semantiikka ja diskurssi. Hän sisällyttää morfologian leksikaaliseen tasoon.

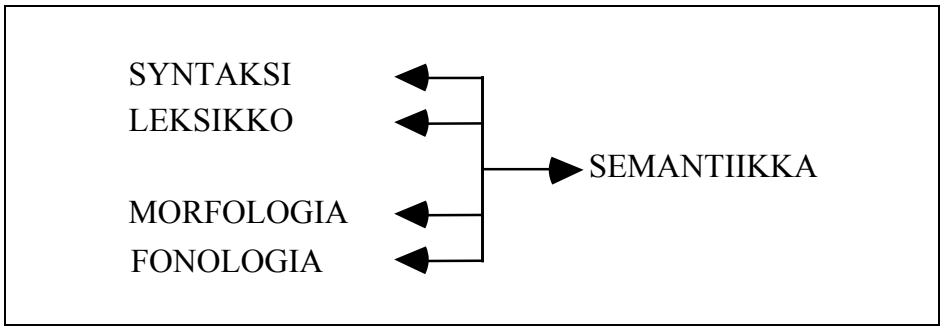
Karlssonin myöhemmässä oppikirjassa (1994) luonnollisen kielen perustavia osajärjestelmiä on viisi: fonologia (ja fonetiikka), morfologia, leksikko (sanakirja), syntaksi ja semantiikka. Pragmatiikasta todetaan, että sen ja semantiikan välinen raja ei ole selvä; pragmatiikka kuitenkin liittyy merkityksen kontekstiriippuvuuteen eli siihen, miten kieltä käytetään eri tilanteissa. Tekstin sidoksisuus ja tekstilingvistiikka, jotka Karlssonin aiemmassa oppikirjassa esitellään Syntaksi-luvussa (1976, s. 218), ovat myöhemmässä kirjassa saaneet oman lukunsa, Teksti ja diskurssi (Karlsson 1994, s. 221).

Seuraavassa osajärjestelmät esitellään Karlssonin (1994) viisijaon mukaisesti.

Kielen osajärjestelmän rakenne koostuu yksiköistä ja näiden välisistä suhteista. Perinteisesti kielen eri osajärjestelmien väliset suhteet on hahmotettu tasojen hierarkiana, joka perustuu koostumissuhteeseen siten, että ylemmän rakennetason yksikkö koostuu alemman tason yksiköistä (kuva 2).

Kuva 2. Kielen viiden osajärjestelmän suhteet perinteisenä tasohierarkiana. Abstraktisuus lisääntyy ylöspäin mentäessä. (Karlsson 1994, s. 15)

Karlsson (1994, s. 16) kuitenkin toteaa edellä kuvatun hierarkian olevan harhaanjohtava semantiikan osalta, sillä semantiikka (tai merkitys) liittyy kaikkiin muihin osajärjestelmiin. Esimerkiksi syntaksin rakenteilla on usein semanttisia erityistehtäviä, morfologiassa päätteillä on tietyt merkityksensä ja niin edelleen. Tasojen suhteet on siis korjattava seuraavankaltaisiksi:



Kuva 3. Kielen viiden osajärjestelmän suhteet korjattuna tasohierarkiana (Karlsson 1994).

Koska kielen eri osajärjestelmien välillä on monenlaisia vuorovaikutussuhteita, ei aina löydy yksiselitteistä luokittelua sille, mihin osajärjestelmään jokin yksittäinen kielen ilmiö kuuluu. Vaikka esimerkiksi yhdyssanojen muodostus kuuluu suomen kielessä luontevimmin morfologiaan, sillä on paljon yhtymäkohtia myös syntaksiin. Erityisesti tämä käy ilmi sellaisissa rajatapauksissa, joissa ei ole selvää, kirjoitetaanko jokin kahden tai useamman sanan yhdistelmä yhteen yhdyssanaksi vai erilleen sanaliitoksi.

Monet tiedonhaun alaan kuuluvat tutkimukset, joissa on käsitelty sanoja tai sanojen välisiä vaikutussuhteita, eivät silti ole olleet luonteeltaan kielitieteellisiä. Sparck Jones (1995) jakaa tiedonhakututkimukset karkeasti kahteen lajiin: toisaalta tilastollisesti ja toisaalta kielitieteellisesti painottuviin tutkimuksiin. Jos jossain tutkimuksessa tarkastellaan haku- ja hakemisto-

sanojen kytkeytymistä toisiinsa sen perusteella, miten sanat esiintyvät toistensa lähellä (proximity) teksteissä, kyseessä on tilastomenetelmiä soveltava tutkimus. Jos vastaavaa ilmiöitä tarkastellaan käyttämällä syntaktisia tai semanttisia määrittelyjä (pyritään tunnistamaan erityisesti sanaliittoja ja yhdyssanoja, esimerkiksi etsitään nominilausekkeen pääsanaa), kyseessä on kielitieteellisesti suuntautunut tutkimus. Eri tutkimushankkeiden sitoutuneisuus omaan suuntaukseensa eli puhdasoppisuus kuitenkin vaihtelee: vaikka joissain hankkeissa olisikin sovellettu vain tilastollisia tai vain kielitieteellisiä menetelmiä, on myös monia tutkimuksia, joissa ensin on sovellettu jompaa kumpaa näistä menettelytavoista ja sen jälkeen korjailtu tai täydennetty tulosta soveltamalla toistakin lähestymistapaa.

Seuraavissa luvuissa esitellään tiedon tallennuksen ja haun tutkimuksia kielen osajärjestelmien mukaisessa järjestyksessä. Katsaus on luonteeltaan esimerkinomainen eli siinä ei pyritä esittelemään kattavasti kaikkia mahdollisia projekteja, joihin kielitiede tai tietokonelingvistiikka jollain tavalla on liittynyt. Laajimmin käsitellään morfologiaan liittyviä tutkimushankkeita, koska ne liittyvät olennaisesti tämän tutkimuksen aihepiiriin. Myös syntaksin alueen tutkimuksia käsitellään suhteellisen laajasti, koska ne näyttävät suuntaa sille, mitkä ovat seuraavat tutkimusalueet, kun suomen kielen morfologiaan liittyvät ongelmat on ratkaistu.

Yksittäisen tutkimushankkeen sijoittaminen vain yhden otsikon alle on usein vaikeaa, koska monissa tiedon tallennuksen ja haun tutkimuksissa on sovellettu useita kielitieteellisiä tai tietokonelingvistisiä menetelmiä, jotka kuuluvat eri osajärjestelmien alueisiin. Esimerkiksi 1980-luvun lopulla rakennetuissa tekoälyjärjestelmissä saatettiin ennen varsinaisten loogisten päättelysääntöjen käyttöä soveltaa niin morfologiaa (syötteenä annettujen merkkijonojen tunnistaminen ja muokkaus sanoiksi), syntaksia (sanojen keskinäisten suhteiden päätteleminen) kuin semantiikkaakin (sanojen merkityksen päättely esimerkiksi sanakirjan avulla). Tällaiset rajatapaukset on pyritty esittelemään luvussa, joka kulloinkin sopii parhaiten kuvattavan tutkimuksen luonteeseen.

4.2 Fonologia ja fonetiikka

Fonetiikka käsittelee puheen tuottamista ja tunnistamista. Siinä analysoidaan puheen akustista rakennetta. Kielen äänteet yksikköinä eli segmentteinä pyritään paljastamaan foneettisen analyysin avulla. Tarkoituksena on

määrittää kielen foneemit eli äännerakenteen kategoriat (äänneperhe). Fonologian tutkimuskohteena taas ovat kielten äännejärjestelmät. Siinä analysoidaan, mikä tehtävä foneettisilla äänne-eroilla on. Tarkoituksena on määritellä, kuuluvatko keskenään erilaiset äänteet eri foneemeihin vai ovatko ne saman foneemin eri edustajia. (Karlsson 1994)

Fonologian ja fonetiikan teorioiden vaikutus tiedon tallennuksen ja haun tutkimukseen on ollut vähäisempi kuin esimerkiksi morfologian tai syntaksin (Warner 1991). Ennen 1990-lukua oli vain muutamia kokeellisia tiedonhakujärjestelmiä, joissa kysely syötettiin hakujärjestelmään puheena. Yleensä tiedonhaku näissäkin projekteissa perustui transkriboituun eli kirjoitetuksi tekstiksi muunnettuun puheeseen eikä äänisignaalien täsmäyttämiseen keskenään (kuten esimerkiksi äänentunnistukseen perustuvissa turvajärjestelmissä). Myöhemmissä tutkimuksissa (kuten Glavitsch & Schäuble 1992; Sparck Jones et al. 1996; Sanderson & Crestani 1998) on kuitenkin tutkittu kyselyn kohdistamista nimenomaan digitoituun äänitietueeseen eikä siitä kirjoitetuksi tekstiksi muutettuun tekstitietokantaan.

Puhutun ja kirjoitetun tekstin erona on muun muassa se, että puhe on jatkuvaa virtaa, jossa sanat eivät välttämättä erotu selkeästi toisistaan - toisin kuin kirjoitetussa tekstissä, jossa eri sanojen välissä on välilyönti tai muu erotin. Puheessa foneemit sitä paitsi voivat ääntyä eri tavoin sen mukaan, mitä foneemeja niiden vieressä on. (Vrt. suomen kielen ilmaus tule tänne, joka tavallisesti äännetään tavalla, jota vastaisi kirjoitusasu tulettänne.) Tämän vuoksi puheen automaattinen tunnistaminen ja muuttaminen tekstiksi on vaikeaa ja puheentunnistusjärjestelmien tuottama teksti siten sisältää enemmän virheellisiä sanoja (kirjoitusvirheitä) kuin ihmisen kirjoittama normaali teksti. (Sanderson & Crestani 1998)

Yksi puheentunnistusjärjestelmien tehokkuusmittari onkin, kuinka suuri osuus niiden tunnistamista sanoista on virheellisiä (word error rate, WER). Puhedokumenttien haussa on olennaista kehittää algoritmeja, jotka sietävät tai ohittavat tunnistusvirheet; ongelma on käytännössä sama kuin käsiteltäessä optisen merkintunnistuksen (optical character recognition, OCR) tuottamaa heikkolaatuista tekstiä (Sanderson & Crestani 1998; Crestani 1999).

Ensimmäisiä foneettis-fonologisia tiedonhakusovelluksia oli kokeellinen HEARSAY-puheentunnistusjärjestelmä (Erman et al. 1980). HEARSAY-II-projektissa kehitettiin tekoälysovellus, joka koostui useista erillisistä, itse-

näisesti toimivista osaohjelmista (knowledge sources). Tarkoituksena oli saada näiden osaohjelmien toiminta koordinoiduksi niin, että ne yhteistyössä kykenivät ongelmanratkaisuun, HEARSAY-projektin tapauksessa tunnistamaan puhetta. Menetelmän toimivuutta testattiin projektissa suppean tiedonhakusovelluksen avulla, jossa viitetietokannasta haettiin tietojenkäsittelyn aihealueeseen kuuluvien julkaisujen referaatteja. Puhuttu hakupyyntö syötettiin HEARSAY-järjestelmään, jonka sitten piti tulkita tämä puhuttu lause kirjoitetuksi. HEARSAY-II:n sanakirjaan sisältyi noin tuhat sanaa ja testikyselyitä oli 23, joissa keskimäärin 7 sanaa. Esimerkkilause oli vaikkapa: "Which abstracts refer to theory of computation?" HEARSAY-järjestelmän toiminta analysoitiin puheentunnistuksen suhteen ja todettiin, että se onnistui tuottamaan lauseelle oikean tulkinnan 90 prosentissa tapauksista. Sen sijaan tiedonhaun onnistumista ei mitattu (tai ainakaan raportoitu).

HEARSAY:n jälkeen on kehitetty useitakin joko puhetta tunnistavia, puhetta tuottavia tai nämä molemmat toiminnot sisältäviä varsinaisia tiedonhakujärjestelmiä. Näiden 1980-luvulla käynnistyneiden hankkeiden tavoitteena on ollut parantaa vammaisten (disabled) kirjastonkäyttäjien tiedonhankintamahdollisuuksia. Tällaisten järjestelmien avulla esimerkiksi näkövammaiset ovat kyenneet käyttämään kirjastoluetteloita (Lange 1993).

Glavitsch ja Schäuble (1992) kehittivät mallin äänidokumenttien hakua varten, aineistonaan radiouutiset. Ensiksi he määrittelivät käsitteen indeksointiyksikkö (indexing feature), jollainen olisi helppo tunnistaa puheentunnistusmenetelmin. He käyttivät tällaisina indeksointiyksikköinä kolmen foneemin yhdistelmää (triphone), joka vastaa tiedonhakututkimuksessa usein käytettyä kolmen merkin yhdistelmää (trigram). Periaatteessa mahdollisista yhdistelmistä valittiin varsinaisiksi indeksointiyksiköiksi vain sellaiset, joita todella voi esiintyä kyseisessä kielessä. Esimerkiksi monet kolmen konsonantin yhdistelmät (vaikkapa xzv) ovat käytännössä mahdottomia. Näin erilaisia indeksointiyksiköitä saatiin kaikkiaan vajaat tuhat kappaletta.

Glavitschin ja Schäublen (1992) mukaan indeksointiyksiköitä voidaan soveltaa yhtä hyvin puhe- kuin tekstidokumenttienkin haussa tai yleensäkin multimediadokumenttien hakemisessa. Mallin toimivuutta tekstidokumenteilla testattiin niin, että vertailukohteina toisaalta olivat haku indeksointi-

yksiköillä ja toisaalta haku ns. perinteisen vektorimallin[1] mukaan. Kun eri menetelmiä vertailtiin saanti-tarkkuus-akselilla, indeksointiyksiköt tuottivat Glavitschin ja Schäublen mukaan jonkin verran parempia tuloksia kuin vertailukohteena ollut vektorimalli.

1990-luvun alussa Sparck Jonesin tutkimusryhmän Video Mail Retrieval (VMR) -projektissa tutkittiin tilastomenetelmiin perustuvaa puheviestien hakua. Projektissa tutkittiin ns. foneettista hakua, joka kohdistui digitoituihin puhedokumentteihin ja niissä esiintyviin sanoihin. Puhedokumenteista pyrittiin automaattisesti havaitsemaan tiettyjä avainsanoja (word spotting), ja näin löydetyistä avainsanoista muodostettiin hakemisto.

Hakujen tuloksellisuutta mitattiin sen mukaan, kuinka hyvin foneettisella haulla löydettiin kaikki - mutta ei enempää kuin - ne dokumentit, joissa hakusanan tiedettiin esiintyvän. Vertailutasona olivat hakutulokset, jotka oli saatu hakemalla vastaavia transkriboituja tekstejä normaalilla sanahaulla. Tuloksena oli, että foneettisen haun tulokset olivat 90 %:isesti samat kuin vastaavan kirjoitettuun tekstiin kohdistetun sanahaun. Foneettisen haun tuloksia huononsivat homofonit eli sananmuodot, jotka puhuttuna kuullostavat samalta, mutta joilla kuitenkin on eri merkitys (esimerkiksi. sanaliitto Hello Kate, jonka loppuosa ääntyy kuten locate-sana). (Sparck Jones et al. 1996)

Käytännössä puheentunnistustekniikat ovat tulleet vasta 1990-luvun lopussa riittävän halvoiksi ja tehokkaiksi soveltuakseen tiedonhakututkimuksen tarpeisiin (Crestani 1999). Yksi osoitus kasvaneesta kiinnostuksesta on, että TREC-6-hankkeeseen perustettiin vuonna 1997 oma osio puhedokumenttien haulle (spoken documents retrieval track, SDR). Siihen osallistui heti ensimmäisenä vuonna 13 tutkimusryhmää, kun samaan aikaan vastaavaan luonnollisen kielen tekniikoita soveltavaan osioon (natural language processing track, NLP) osallistui vain kaksi ryhmää (Sparck Jones 2000).

---

[1] Vertailukohteena olleessa vektorimenetelmässä käytettiin sulkusanalistaa ja Porterin (1980) karsinta-algoritmia. Hakemistosanojen painoarvot laskettiin kaavalla *tf*idf* (*term frequency times inverse document frequency* eli sanan esiintymistiheys dokumentissa kertaa sen käänteinen esiintymistiheys koko kokoelmassa). Kutakin dokumenttia (tai vastaavasti kyselyä) kuvasi vektori, joka sisälsi painotetut hakemistosanat (hakusanat). Dokumentti- ja kyselyvektoreiden täsmäyttämisessä käytettiin kosinikaavaa. (Glavitsch ja Schäuble 1992)

Tutkimuksen edistymisestä huolimatta ovat tämänhetkiset puheen tunnistus- ja hakujärjestelmät vielä rajoittuneita. Usein ne perustuvat rajalliseen sanastoon tai ne on viritettävä yhden tietyn henkilön puhetta varten. Lisäksi tunnistustarkkuus kärsii, kun puhe on normaalia jatkuvaa virtaa eikä selkeitä erillisiä sanoja. Vaikka erilaisia prototyyppejä on kehitetty, ei normaalia, rajoittamatonta puhetta ymmärtäviä puheentunnistusjärjestelmiä vielä tällä hetkellä ole yleisesti saatavilla. Tosin markkinoilla olevat muutamat kaupalliset puheentunnistusjärjestelmät suoriutuvat tehtävästään varsin hyvin, kunhan edellä mainitut rajoitukset otetaan huomioon. (Crestani 1999)

Puhedokumenttien hakujärjestelmien kaupallinen läpimurto on kuitenkin lähitulevaisuudessa odotettavissa, koska alan tutkimus on saanut lisäpontta matkapuhelimien yleistyttyä tavanomaiseksi yhteydenpitovälineeksi länsimaissa - tarvetta ja potentiaalista käyttäjäkuntaa puheentunnistusta käyttäville palveluille, kuten matkalippujen tilaamiselle tai uutisten kuuntelulle matkapuhelimitse, alkaa olla riittävästi (Crestani 1999).

4.3 Morfologia

Morfologia eli muoto-oppi tutkii morfeemien rakennetta ja käyttäytymistä. Tiedonhaussa morfologia liittyy hakusanojen ja hakemistosanojen esitysmuotoon. Perinteisissä hakujärjestelmissähän dokumenteissa esiintyneet sanat tallennetaan hakemistoon siinä muodossa kuin ne tekstissä esiintyvät. Käytännössä hakemistoon ei siis tallenneta sanoja (lekseemejä), vaan sanojen esiintymiä (sananmuotoja). Hakuvaiheessa käyttäjän on otettava tämä huomioon ja käytettävä merkkijonokaavioita, jotka kattavat hakusanan kaikki mahdolliset esiintymät. Tiedonhaun yksi keskeinen ongelma onkin, miten merkitykseltään lähekkäiset, mutta merkkijonoina erilaiset sanat ja sananmuodot saataisiin yhdistettyä (conflation) toisiinsa.

Eri sananmuotojen korvaamista yhdellä vakiomuodolla (sanan vartalo, perusmuoto tms.) on perusteltu muun muassa seuraavilla hyödyillä:

1. muistitilan säästäminen
2. tulosjoukon saannin parantaminen
3. dokumentin painokertoimien täsmällisempi laskenta

Nykyään on niin teknisesti kuin taloudellisestikin helppoa hankkia tiedonhakujärjestelmään niin paljon tallennusmuistia kuin tarvitaan. Aiemmin tämä ei ollut yhtä yksinkertaista, vaan tietokoneiden muistit olivat nykyistä huomattavasti pienempiä ja hinnaltaan kalliimpia. Tällöin oli välttämätöntä saada muistitilan tarve niin pieneksi kuin mahdollista. Yksi säästökeino oli pienentää hakujärjestelmän hakemiston kokoa siten, että eri sananmuotojen sijasta hakemistoon tallennettiin vain niiden yhteinen vartalo (Bell & Jones 1979; Jones & Bell 1986).

Toinen syy eli saannin parantaminen on ollut kimmokkeena lukuisissa tutkimuksissa. Niiden lähtökohtana on ollut, että käyttäjä antaa hakusanan jossain vakiomuodossa, ja hakujärjestelmä sitten laajentaa kyselyä automaattisesti hakusanan varianteilla. Oletuksena on, että merkitykseltään läheiset ilmaukset olisivat myös kirjoitusasultaan lähellä toisiaan. Tällöin hakujärjestelmä korvaa käyttäjän antaman yksittäisen hakusanan (merkkijonovakion) hakusanalla tai -sanoilla (merkkijonokaavioilla), jotka kattavat kaikki hakusanalle läheiset morfologiset variantit, lähinnä taivutusmuodot ja johdokset. (Porter 1980, Ulmschneider & Doszkocs 1983, Harman 1987; 1991; Paice 1990)

Kolmas syy eli painokertoimien laskenta on rinnakkainen edelliselle vaihtoehdolle, siinä vain haun saantia parannetaan epäsuoremmin. Jos nimittäin jokaisen dokumentissa esiintyneen sananmuodon eli käytännössä merkkijonon taajuus lasketaan erikseen, ei saada totuudenmukaista kuvaa itse sanan esiintymistaajuudesta. Kun sanan eri esiintymien taajuudet summataan yhteen ja sanalle annetaan näin saatu kokonaiskerroin, tällaisten painokertointen oletetaan kuvaavan dokumentin sisältöä todenmukaisemmin (Salton 1989, s. 304).

4.3.1 Keinoja sananmuotojen yhdistämiseksi

Alkeellisin keino mahdollistaa se, että yhdellä hakusanalla voidaan palauttaa erilaisia merkkijonoja, on käyttää **jokerimerkkejä**. Jokerimerkki voi olla esimerkiksi symboli "*", jolla korvataan mielivaltainen määrä merkkejä sananmuodon lopusta (mahdollisesti myös sananmuodon alusta tai keskeltä, jos se hakujärjestelmässä sallitaan). Jokaista hakusanan sananmuotoa ei tarvitse erikseen kirjoittaa kyselyyn, vaan jokerimerkillä varustettu hakusana poimii hakemistosta kaikki samalla merkkijonolla alkavat sananmuodot (esimerkiksi *demokra** -> demokratia, demokratian, demokraattinen, demo-

kraatti, demokraattia jne.). Merkin **peittäminen** (masking) taas on menetelmä, jolla sallitaan sanan sisällä esiintyvät vaihtelut. Jos "**!**" symboloi yhden merkin peittämistä, niin hakusana *nitr!!tti* palauttaa sekä sanan nitriitti että nitraatti.

Jokeri- ja peittomerkit ovat apukeinoja tilanteessa, jossa hakija itse huolehtii hakusanan katkaisusta. Seuraava askel on, että hakujärjestelmä hoitaa hakusanan katkaisemisen. Englannin kielessä tämä on toteutettu ns. **karsinta-algoritmeilla** (stemming algorithm, suffixing, suffix stripping), jotka poistavat sananmuotojen lopusta taivutuspäätteet ja johtimet. Yleensä karsinta-algoritmit siis poistavat vain pääteaineksia, jotkut käsittelevät myös etuliitteitä (Paice 1990).

Koska karsinnan tarkoituksena on yhdistää erilaiset merkkijonot kyselyssä, karsinnan tuloksena saatava muoto ei välttämättä ole sanan todellinen vartalo tai kanta. Menetelmän laatijan pyrkimyksenä voi esimerkiksi olla saada karsintamenetelmästä mahdollisimman johdonmukainen tai tehokas - olkoonkin, että se joissain tapauksissa tuottaisi kieliopilliselta kannalta virheellisen vartalon (Paice 1990; Salton 1989, s. 382).

Yksinkertaisin karsintamenetelmä on, että kaikki sananmuodot katkaistaan tietynmittaisiksi merkkijonoiksi, esimerkiksi kahdeksan merkin mittaisiksi. Tällainen menetelmä kuitenkin tuottaa aivan satunnaisia tuloksia, koska luonnollisen kielen sanat ovat erimittaisia - se toimisi toivotulla tavalla vain kielessä, jossa kaikki sanat ovat yhtä pitkiä. Tätä menetelmää ei ole sovellettu missään kokeellisessakaan tiedonhakujärjestelmässä kuin perustasona, johon muita karsintamenetelmiä verrataan: kelvollisen karsinta-algoritmin pitää olla parempi kuin pelkän sanojen määrämittaiseksi katkaisun, jotta sitä ylipäätään kannattaisi käyttää (Lennon et al. 1981; Paice 1994).

Varsinaisissa karsinta-algoritmeissa hyödynnetään tietoja kielen morfologisista ominaisuuksista. Esimerkiksi englannin kielessä voidaan erottaa noin 75 etuliitettä ja noin 250 päätettä (Salton 1989, s. 380). Tyypillisesti karsinta-algoritmit sisältävät luettelon päätteistä ja johtimista sekä säännöt, joiden mukaan tällaiset pääteainekset karsitaan pois niin, että jäljelle jää sanan vartalo (stem). Päätteiden karsimisen lisäksi säännöillä pyritään myös estämään epäkelpojen vartaloiden syntyminen. Voidaan esimerkiksi määritellä, että jäljellejäävän vartalon on oltava vähintään neljä merkkiä pitkä,

jotta päätteen saa karsia sananmuodon lopusta. On tarkoituksenmukaista poistaa ing-pääte interesting-sanan lopusta, muttei king-sanasta.

Karsinta-algoritmit ovat yleensä sääntöpohjaisia eivätkä siten sisällä varsinaista sanakirjaa. Ne voivat silti sisältää sulkusanalistan, jolla estetään haun kannalta tarpeettomien sanojen, kuten prepositioiden ja konjunktioiden, käsittely (Ulmschneider & Doszkocs 1983). Täysin sääntöpohjaisten karsinta-algoritmien puutteena kuitenkin on, että niillä voidaan yhdistää toisiinsa vain säännöllisesti taipuvat sanat, joiden sananmuodoilla on yhteinen vartalo. Sen sijaan epäsäännöllisesti taipuvien sanojen eri taivutusmuodot, kuten freeze ja frozen tai mouse ja mice, jäävät toisistaan erilleen. Tämän vuoksi kehittyneimmissä karsinta-algoritmeissa voi olla vielä poikkeussanakirja, jonka perusteella tuotetaan normalisoidut vartalot, ts. epäsäännöllisesti taipuvien sananmuotojen päätteitä ei karsita lainkaan, vaan nämä sananmuodot korvataan suoraan sanakirjasta löytyvällä vartalolla (Glassco 1993).

Yksi yleisimmin käytetyistä englannin kielen karsinta-algoritmeista on Porterin (1980) kehittämä. Se on ns. yleisalgoritmi, joka ei ole sidottu mihinkään erityisalueen tekstiin. Joissain tutkimuksissa on kehitetty myös erityisalgoritmeja, joissa otetaan huomioon tekstin erityisala ja sen sanasto. Tällainen on muun muassa CITE/CATLINE-näyttöluetteloon kehitetty algoritmi, joka on viritetty lääketieteellisen tekstin käsittelyyn (Ulmschneider & Doszkocs 1983). Siihen on muun muassa lisätty kreikasta ja latinasta lähtöisin olevia, lääketieteellisessä kielenkäytössä esiintyviä päätteitä (kuten -sis, -ses).

4.3.2 Englannin karsinta-algoritmien kritiikki - onko karsinnalla väliä?

Karsinta-algoritmien vaikutusta tiedonhakuun on selvittänyt muun muassa Harman (1987; 1991), joka vertaili kolmea englannin kielen yleisalgoritmia. Sysäyksenä hänen tutkimukselleen oli se seikka, että tiedon tallennuksen ja haun (lähinnä tilastollisten menetelmien) tutkijat eivät juurikaan maininneet raporteissaan, oliko hakusanat normalisoitu jotain karsinta-algoritmia käyttäen, saati että käytettyä algoritmia olisi kuvattu. Eri tutkimushankkeissa saatujen tulosten vertaileminen keskenään oli hankalaa. Eri menetelmien välisiä eroja ja niiden vaikutusta hakujen tuloksiin ei juuri ollut tutkittu (paitsi Lennon et al. 1981), vaikka karsinta-algoritmien soveltaminen oli tiedonhakututkimuksissa yleinen käytäntö.

Harman (1987; 1991) vertaili Lovinsin, Porterin ja S-algoritmia keskenään. Näistä yksinkertaisin on S-algoritmi, joka poistaa sananmuodon lopusta monikon tunnuksen. Lovinsin algoritmi puolestaan sisältää 260 päätteen sekä poikkeussanojen luettelon. Se poistaa sananmuodon lopusta pisimmän mahdollisen päätteen, edellyttäen että jäljelle jää riittävän pitkä sanavartalo (vähintään kaksi merkkiä). Porterin (1980) algoritmissa taas pääteaineksia karsitaan vaiheittain pois niin, että käsittelyn eri vaiheissa joko poistetaan pääte tai tehdään sanavartalolle jokin muunnos. Tämä algoritmi tunnistaa noin 60 eri päätettä.

Harman teki vertailut osittaistäsmäytykseen perustuvassa IRX-hakujärjestelmässä. IRX-järjestelmän hakemisto oli sama kaikissa testeissä eli sitä ei käsitelty eri karsinta-algoritmeilla, vaan ainoastaan kyselyt. Testikyselyjen sanat syötettiin vertailtaville karsinta-algoritmeille ja niiden tuottamat vartalot lisättiin kyselyyn. IRX:n rankkeerausmenetelmää muokattiin siten, että kun se asetti dokumentteja paremmuusjärjestykseen, se käsitteli sanan eri variantteja yhtenä kokonaisuutena, sen sijaan että olisi pitänyt eri sananmuodot erillisinä. (Harman 1991)

Harman (1987; 1991) päätyi tulokseen, että karsinta-algoritmien käyttö ei paranna tiedonhaun tuloksia. Kun kaikista saanti- ja tarkkuusarvoista laskettiin keskiarvot, nämä olivat suunnilleen samat sekä alkuperäisillä, käsittelemättömillä hakusanoilla että karsituilla hakusanoilla saaduilla tulosjoukoilla. Harman kuitenkin kertoo, että kun tulosjoukkoja vertailtiin keskenään, alkuperäisillä karsimattomilla hakusanoilla saaduissa tulosjoukoissa oli dokumentteja, joita ei karsituilla hakusanoilla saaduissa tulosjoukoissa ollut, ja päinvastoin. Ongelmana vain oli, että karsinnan tuloksena saadut vartalot tuntuivat toimivan satunnaisella tavalla: toiset tuottivat hyödyllisiä osumia, kun taas toisten vartaloiden avulla saatiin ei-toivottuja dokumentteja.

Harmanin käyttämässä IRX-testijärjestelmässä hakusanoille laskettiin painoarvot niiden esiintymistiheyden mukaan. Karsinta-algoritmien tuottamat hyödyttömät vartalot saattoivat lisätä epärelevantin dokumentin painoarvoa ja nostaa sen virheellisesti relevanttien dokumenttien yläpuolelle paremmuusjärjestyksessä. Saannin ja tarkkuuden keskiarvot pysyivät samoina, käytettiinpä karsinta-algoritmeja tai ei, koska joidenkin kyselyjen tulokset paranivat ja toisten huononivat verrattuna lähtötilanteeseen.

Harmanin (1987) havaitsema ongelma siis oli, että muodollisin perustein toimivien karsinta-algoritmien tuottamien vartaloiden hyödyllisyyttä tai hyödyttömyyttä ei voi päätellä vastaavasti muodollisin perustein. Harman tekikin toisen tutkimuksen (1988), jossa vartalot ensin esitettiin hakijalle, joka valitsi näistä sopiviksi katsomansa, ja vain nämä hyväksytyt vartalot lisättiin kyselyyn. Harmanin mukaan näillä valikoiduilla vartaloilla saatuun tulosjoukkoon tuli vähemmän ei-toivottuja dokumentteja kuin silloin, kun tällaista ennakkovalikointia ei tehty. Manuaalisella valikoinnilla oli siis mahdollista paikata karsinta-algoritmien tuotosta ja parantaa haun saantia ilman vastaavaa tarkkuuden romahtamista. Tämän perusteella Harman (1987; 1991) päätyi suosittelemaan menetelmää, jossa karsinta-algoritmin käyttö tai käyttämättä jättäminen olisi hakijan valittavissa. Hakujärjestelmän oletusarvona olisi, että hakusanat karsitaan automaattisesti, mutta tästä poikettaisiin, jos tulosjoukkoon tulee liikaa ei-toivottuja dokumentteja. Tällöin kysely suoritettaisiin uudelleen niin, että hakusanoja ei karsittaisi karsinta-algorimilla.

Valitettavasti Harmanin kokeilu oli suppea eikä siten anna riittävästi takeita manuaalisen valikoinnin hyödyllisyydestä. Esimerkiksi Krovetz (1993) testasi valikointia omissa testikokoelmissaan ja havaitsi, että valikoinnin vaikutus vaihteli eri kokoelmissa: joissain testikokoelmissa se paransi ja toisissa taas huononsi hakutuloksia. Toinen Harmanin ehdottaman menetelmän ongelma on, ettei se onnistu hakujärjestelmässä, jossa sananmuodot on karsittu jo tallennusvaiheessa hakemiston muistitilan säästämiseksi - kun karsinnan aste on lyöty lukkoon jo tallennettaessa, sitä ei voi enää hakuvaiheessa säädellä (Keen 1991).

Mielenkiintoista on, että vaikka Harman (1987; 1991) toteaakin lisätutkimukset tarpeellisiksi, hän tutkimuksensa yhteenvedossa tekee varsin pitkällevietyjä johtopäätöksiä ja kyseenalaistaa karsinta-algoritmien hyödyllisyyden saamiensa yleisten keskiarvolukujen perusteella. Yksittäisissä hauissahan karsinta-algoritmeilla oli selvästi vaikutusta haun tuloksiin. Sitä paitsi Harmanin tarkastelemat tulosjoukot eivät syntyneet pelkästään karsinnan tuloksena, vaan tulosjoukkojen koostumukseen vaikuttivat karsinta-algoritmit ja tilastomenetelmät yhdessä. IRX-järjestelmän rankkeeraustavan mahdollista osuutta tuloksiin Harman ei pohdi lainkaan.

Harmanin tutkimustulokset kenties olisivat olleet toisenlaiset, mikäli tulosjoukkoja olisi analysoitu yksityiskohtaisemmin ja tutkittu, minkätyyppisillä

hakusanoilla kulloinkin oli haettu. Onhan mahdollista, että tarkempi jaottelu olisi paljastanut sana- tai päätetyyppejä, joissa karsinnan tulokset eivät ole niin satunnaisia kuin Harmanin testikyselyjen keskiarvot näyttävät. Tosin Keen (1991) totesi oman tutkimuksensa perusteella, että tällaisia varmoja karsittavia ei ole. Ongelmana on, että johtimet eri sanoissa vaikuttavat eri tavoin sanan merkitykseen. Jossain tapauksessa johdin muuttaa alkuperäisen sanan merkitystä vain vähän (censor + ing -> censoring), mutta toisessa tuottaa semanttisesti sangen kaukaisen johdoksen (key + ing -> keying). Harmanin ja Keenin tutkimuksissa karsinta-algoritmit siis toimivat jokseenkin sattumanvaraisesti: ne saattoivat tuottaa yhtä hyvin alkuperäistä kyselyä parempia kuin sitä huonompia tuloksia. Lopputulos riippui siitä, millaisia johtimia sanalla sattui olemaan ja miten kyseiset johtimet vaikuttivat sanan merkitykseen.

Harmanin (1987; 1991) vertailu ei ole ainoa laatuaan. Sitä on edellä käsitelty laajasti siksi, että Harmanin artikkelin (1991) "How effective is suffixing?" painoarvo on informaatiotutkijoiden keskuudessa ollut suuri, onhan Harman kansainvälisesti tunnettu tutkija. Ennen Harmania myös Lennon et al. (1981) olivat vertailleet eri karsinta-algoritmeja ja todenneet niiden tuottavan varsin vähän eroja tulosjoukkojen välille. Heidän tutkimuksessaan vertailtiin seitsemää eri algoritmia. Tutkimusaineistona oli noin 1 400 kappaletta viitetietueiden otsikkokenttiä ja 225 kyselyä. Näissä esiintyneet sananmuodot karsittiin kullakin tutkittavalla algoritmilla. Algoritmeista tutkittiin, kuinka tehokkaasti ne supistavat hakemiston kokoa (eli vähentävät hakemiston merkkijonojen määrää) ja toisaalta lisäävät haun tehokkuutta. Haun tehokkuus laskettiin sekä saanti- että tarkkuusarvot huomioonottavan laskukaavan avulla. Ennakko-oletuksena oli, että taivutuspäätteitä ja johtimia rankasti karsivat algoritmit ovat myös saanniltaan parhaita. Tulokset kuitenkin osoittivat, että karsinnan laajuudella ja haun saannilla ei ollut merkittävää yhteyttä.

Toisaalta Lennon et al. (1981) totesivat, että mikään karsinta-algoritmeista ei ollut tehokkuudeltaan huonompi kuin karsimatta jättäminen. Päinvastoin, useat algoritmit olivat tehokkuusluvuiltaan huomattavasti tätä nollavaihtoehtoa parempia. Tämän vertailun lopputulokseksi saatiin, että vaikka mikään algoritmeista ei osoittautunut toisia merkittävästi paremmaksi, karsiminen sinänsä voi parantaa tiedonhaun tehokkuutta: karsinta-algoritmit voivat tuottaa parempia hakutuloksia kuin karsimatta jättäminen - ainakaan eivät huonompia.

Myös Keen (1991) oli sitä mieltä, että karsinta ei vaikuta kovin paljon hakutuloksiin, mutta yleisesti ottaen tuottaa parempia tuloksia kuin karsimatta jättäminen. Ensinnäkin, Keenin mukaan Cranfield 2 -tutkimuksessa aikoinaan käytetyt erilaiset karsinta-algoritmit tuottivat keskenään hyvin samankaltaisia tuloksia, vaikka algoritmien toimintaperiaatteet lingvistisessä mielessä saattoivat huomattavastikin poiketa toisistaan. Toiseksi, Keenin omassakaan tutkimuksessa eri karsintatapojen välille ei löytynyt merkittäviä eroja. Keen kuitenkin huomauttaa, että Harmanin (1991) vertailuissa olisi kannattanut ottaa huomioon erojen johdonmukaisuus, vaikka tilastollisesti merkitseviä eroavuuksia ei löydettykään. Kun Harmanin tutkimuksessa oli 198 pareittaista vertailua, jossa karsimattoman ja karsitun hakusanan tuottamia tuloksia verrattiin, 131 tapauksessa jokin karsintamenetelmistä tuotti paremman tuloksen, 47 tapauksessa tulokset olivat yhtä hyvät ja vain 20 tapauksessa karsimaton hakusana tuotti paremman tuloksen kuin karsittu. Joten Harmanin johtopäätös, että pääteainesten karsinta ei vaikuta hakutuloksiin, ei täysin pidä paikkaansa - oikeastaan Harmanin vertailuissa saatiin samansuuntainen tulos kuin Lennonin et al. (1981) vertailuissa.

On myös tutkimuksia, joissa karsinta-algoritmi on selkästi todettu hyödylliseksi englanninkielisen aineiston käsittelyssä. Esimerkki tästä on Okapi-näyttöluettelo, jossa kokeiltiin kahta eriasteista karsintamenetelmää, ns. heikkoa ja vahvaa karsintaa (Walker 1988). Heikossa karsinnassa (weak stemming) sanojen lopusta poistettiin pelkästään monikon tunnus -s sekä -ing ja -ed-päätteet. Vahva karsinta (strong stemming) puolestaan toteutettiin Porterin (1980) algoritmin mukaisesti, sovitettuna brittiläisen ja amerikanenglannin eroihin. Walker (1988) totesi, että heikko karsinta lähes poikkeuksetta paransi haun tuloksia, ts. saanti kasvoi ilman, että tarkkuus olisi juurikaan huonontunut. Sen sijaan vahva karsinta muutti haun tulosta monesti ei-toivottuun suuntaan, ts. vaikka saanti paranikin, tarkkuus huononi selvästi. Keen (1991) kuitenkin epäili omien vertailujensa perusteella, että -ing ja -ed-päätteiden poistaminen ei ole niin hyödyllistä kuin Okapi-tutkimuksessa esitettiin.

Kuten edellä on käynyt ilmi, englanninkielisten haku- ja hakemistosanojen karsinnan hyödyllisyydestä tai hyödyttömyydestä ei vielä 1990-luvun alkupuolellakaan ollut selvää käsitystä. Ilmeisesti tiedonhakututkijat ovat intuitiivisesti pitäneet karsintaa kuitenkin hyödyllisenä, koska sitä on yleisesti sovellettu. Esimerkkinä voidaan mainita vaikkapa TREC: käytännöllisesti katsoen kaikissa TREC-hankkeen kakkosvaiheen projekteissa oli käytetty

jonkinlaista karsintamenetelmää (Sparck Jones 1995). Karsinta-algoritmien käyttökelpoisuus tuli lopulta todistetuksi, kun Hull (1996) vertaili niitä syvällisemmin kuin aiemmissa tutkimuksissa muun muassa tilastollisten merkitsevyystestien avulla. Hänen analyysin perusteella karsinta osoittautui hyödylliseksi. (Hullin tutkimuksen tarkempi selostus seuraavassa luvussa.)

4.3.3 Karsinta-algoritmien edelleenkehittäminen

Paice (1994) toteaa, että karsinta-algoritmien vertailu pelkästään hakutulosten perusteella ei anna riittävän tarkkaa tietoa siitä, mihin tekijöihin niiden vaikutus perustuu ja mistä virheet johtuvat. Kun karsinta-algoritmin toiminnasta saadaan tietoa vain epäsuorasti sen tuottamien tulosjoukkojen perusteella, algoritmin hienosäätö on vaikeaa. Sitä paitsi karsinta-algoritmeja sovelletaan muuallakin kuin tiedonhaussa, esimerkiksi luonnollisen kielen käyttöliittymissä. Algoritmeja tulisikin arvioida ja kehittää itsenäisesti eikä vain sen perusteella, millaisia tulosjoukkoja ne tuottavat - etenkin, kun hakutuloksia vertailemalla ei eri algoritmien välille oikein ole löydetty eroja (vrt. Harman 1987; 1991; Lennon 1981, Keen 1991).

Esimerkiksi Harman käsitteli karsinta-algoritmeja kokonaisuuksina, eräänlaisina jakamattomina mustina laatikkoina, joiden toimintaa ei voi hienosäätää. Algoritmia joko sovelletaan tai sitten ei. Todellisuudessa karsinta-algoritmit koostuvat joukosta päätteitä ja sääntöjä, joita on mahdollista muokata. Paice (1990) toteaa, että sekä Porter että Lovins artikkeleissaan ensinnäkin kuvaavat karsinta-algoritmiensa yleisperiaatteet ja toiseksi esittävät sääntökokoelman, jonka perusteella algoritmia käytännössä sovelletaan. Nämä algoritmit eivät siis ole "kiveen hakattuja", vaan sääntökokoelmaa on täysin mahdollista muokata ja säätää - myös tietyn sovelluksen erityistarpeiden mukaan.

Karsinnan lähtökohtana on, että sen pitäisi ryhmitellä samanmerkityksiset sananmuodot samaan joukkoon ja erottaa erimerkityksiset sananmuodot toisistaan. Tällöin oletetaan, että samaan käsitteeseen[2] viittaavat ilmaukset

[2] Paicen käyttämä termi on "concept" - tarkemmin sanoen kyseessä on paremminkin johdosperhe eli sana ja sen kantasana ja johdokset, jotka ovat merkkijonoina lähellä toisiaan; samaan käsitteeseen viittaavat ilmauksethan voivat olla hyvin erilaisia merkkijonoja, kuten ilta ja ehtoo. Joka tapauksessa Paice viittaa lekseemiä eli sanaa laajempaan ilmaukseen - karsinta ei kohdistu vain sanan taivutuspäätteisiin vaan myös johtimiin.

ovat myös merkkijonoina hyvin samankaltaisia. Tällainen suhde on tyypillisesti esimerkiksi kantasanan ja sen johdosten välillä.

Karsinnassa voi tapahtua kahdentyyppisiä virheitä. **Alikarsinnassa** (understemming) sananmuodot, jotka viittaavat samaan johdosperheeseen, eivät päädy samaan joukkoon; niiden vartalot eivät karsinnan jälkeen ole samat, vaikka näin pitäisi olla. **Ylikarsinnassa** (overstemming) taas sananmuodoista saadaan sama vartalo, vaikka todellisuudessa on kyse eri käsitteisiin viittaavista sananmuodoista. Karsinta-algoritmin kehittämisessä joudutaankin tasapainoilemaan näiden kahden virhetyypin välillä: **kevyt karsinta** (light stemmer) poistaa vain varmaksi tiedetyt pääteainekset, mutta ongelmaksi jää alikarsinta; **järeä karsinta** (heavy stemmer) poistaa suoraviivaisesti kaikenlaiset päätteet, jolloin niitä saattaa pyyhkiytyä pois liikaakin ja seurauksena on ylikarsinta. (Paice 1994)

Paicen (1994) tutkimuksessa käytiin ensin manuaalisesti läpi otsikoista ja tiivistelmistä koostuva tutkimusaineisto. Aineistossa esiintyvät sananmuodot ryhmiteltiin semanttisin perustein eri ryhmiin. Tämän jälkeen tutkittiin, kuinka hyvin karsinta-algoritmit pystyivät automaattisesti ryhmittelemään nämä sananmuodot oikeisiin ryhmiin. Algoritmien onnistumista mitattiin yli- ja alikarsinnan perusteella. Vertailtavina olivat Porterin, Lovinsin ja Paicen/Huskin (Paice 1990) algoritmit. Lisäksi vertailujen perustasona olivat sananmuotojen katkaisu määrämittaisiksi merkkijonoiksi; Paice kokeili katkaisua 4, 5, 6, 7 ja 8 merkin mittaisiksi merkkijonoiksi.

Vertailujen tuloksena Paice totesi, että Porterin algoritmi on kevyt karsinta-algoritmi ja että Lovinsin algoritmi on varsin lähellä sitä, vaikkakin hiukan järeämpi. Paicen ja Huskin algoritmi oli selvästi järein näistä kolmesta. Vertailussa havaittiin myös, että eri tutkimusaineistolla (lähdetekstillä) Paicen mittarit tuottivat selvästi erilaisia arvoja. Tämän perusteella Paice sanoo, että eri algoritmeja ei ehkä kannattaisikaan asettaa suoraan paremmusjärjestykseen, koska se saattaa vaihdella tapauksittain: joissain tekstityypeissä ylikarsinta on pienempi haitta kuin alikarsinta, toisissa taas päinvastoin. Sen sijaan Paicen mittareita tulisi hyödyntää karsinta-algoritmien jatkokehityksessä. Jos esimerkiksi Porterin algoritmia pidetään jonkin sovelluksen tarpeisiin liian kevyenä, sitä voidaan järeyttää esimerkiksi lisäämällä siihen uusia sääntöjä tai muokkaamalla entisiä. (Paice 1994)

Paicen lisäksi myös Krovetz (1993) on esittänyt uusia näkökohtia karsintamenetelmien kehittämiseksi. Krovetz kritisoi sitä, että englannin kielen algoritmeissa ei juuri ole otettu huomioon sanojen merkitystä eikä tehty eroa taivutuspäätteiden ja johdinten välillä. Useimmat karsinta-algoritmit toimivat sääntöpohjaisesti, ilman sanakirjaa, mikä voi johtaa vääriin tulkintoihin. Tämän korjaamiseksi Krovetz kehitti omia karsinta- tai oikeastaan perusmuotoistamismenetelmiä, joissa hyödynnettiin sanakirjaa.

Krovetz vertaili neljää karsinta-algoritmia karsimatta jättämisen kanssa. Ensimmäinen oli alkuperäinen Porterin (1980) algoritmi ja toinen Krovetzin korjaama Porterin algoritmi. Kolmas oli Krovetzin taivutusmuotoalgoritmi (inflectional stemmer), jossa monikkomuodot palautettiin yksikköön, imperfektimuodot preesensiin ja poistettiin ing-pääte. Neljäntenä oli Krovetzin kehittämä johdosalgoritmi, jossa poistettiin englannin kielen viisitoista yleisintä johdinta. (Krovetz 1993)

Jokaisessa Krovetzin (1993) algoritmissa sovellettiin sanakirjaa. Lähtökohtana oli, että jos käsiteltävä sananmuoto löytyi sellaisenaan sanakirjasta, kyseessä oli itsenäinen lekseemi, jota ei enää tarvinnut käsitellä enempää. Jos taas sananmuoto ei löytynyt sanakirjasta, tehtiin uusi karsintakierros. Joka kerta, kun pääteaineksia poistettiin, näin saatua sananmuotoa verrattiin sanakirjaan, kunnes sieltä löytyi vastaava lekseemi. Tavoitteena oli, että käsittelyn tuloksena saataisiin paremminkin sana (ts. sanan perusmuoto) kuin vartalo.

Aineistona oli neljä erilaista tekstikokoelmaa. Vertailut tehtiin laskemalla keskimääräinen tarkkuus vakioiduilla saantitasoilla. Vertaillut saantitasot olivat 25, 50 ja 75 prosenttia. Porterin alkuperäinen algoritmi oli useimmissa kokoelmissa ja saantitasoilla parempi eli tuotti korkeammat tarkkuusarvot kuin karsimatta jättäminen. Korjattu Porterin algoritmi oli aina parempi kuin karsimatta jättäminen, mutta joissain kokoelmissa ja saantitasoilla sen tarkkuus kuitenkin saattoi jäädä huonommaksi kuin alkuperäisen Porterin algoritmin. Taivutusmuotoalgoritmi oli - yhden kokoelman yhtä saantitasoa lukuunottamatta - aina parempi kuin karsimatta jättäminen, mutta toisaalta sen tarkkuustaso jäi useimmiten alle sekä alkuperäisen Porterin että korjatun Porterin algoritmin tarkkuustason. Parhaimmat tulokset saatiin Krovetzin johdosalgorimilla. Se oli aina parempi kuin karsimatta jättäminen ja kolmessa kokoelmassa neljästä myös parempi kuin muut algoritmit. (Krovetz 1993)

Krovetz totesi myös seuraavat seikat:

1. Mitä lyhyempiä dokumentit ovat, sitä selvemmin päätteiden karsinta (perusmuotoistaminen) parantaa hakutuloksia
2. Karsinta tuottaa sitä parempia tuloksia, mitä korkeammalla saantitasolla ollaan
3. Erityisesti johdinten, siis johdosten, merkitys kasvaa, kun saanti on alhainen eli kun hakija painottaa tarkkuutta ja hänelle riittää vain muutaman dokumentin läpikäynti.

Krovetzin (1993) vertailut osoittivat, että morfologisesti hienosäätöisempi pääteainesten poistaminen voi parantaa hakutuloksia. Krovetzin menetelmän etuna on myös se, että kokonaisia sanoja pystytään jatkokäsittelemään paremmin kuin vartaloita: esimerkiksi monitulkintaisuuksien karsinta on mahdollista. Järeällä karsinta-algoritmilla tuotettuja vartaloita ei voi enää yksiselitteistää, koska karsinnan jälkeen ei tiedetä, mistä sanasta on kyse. Toisaalta Krovetzin menetelmän ja yleensäkin sanakirjan käyttöön perustuvien menetelmien ongelmana on, että on keksittävä jokin erillisratkaisu sellaisille sananmuodoille, joita sanakirjasta ei löydy ja joita algoritmi siten ei pysty käsittelemään.

Tähän mennessä perusteellisimman vertailun eri katkaisualgoritmien välillä on tehnyt Hull (1996). Hän käytti tutkimuksessaan Xeroxin kehittämää algoritmia, jota oli mahdollista säätää käsittelemään joko pelkästään taivutuspäätteitä tai sekä taivutuspäätteitä että johtimia. Aineistona oli TREC:ssä käytetty Wall Street Journalin noin 180 000 artikkelia sisältänyt osakokoelma ja 200 hakupyynnön joukko. Hakujärjestelmänä oli vektorimalliin perustuva SMART. Kyselyt olivat joko ns. pitkiä kyselyitä, jotka olivat TREC-projektin alkuperäisiä, pitkiä hakupyynnön aihepiirin ja relevanssin määrittelyjä, tai sitten ns. lyhyitä kyselyjä, joihin oli pyritty tiivistämään edellä mainituissa määrittelyissä esiintyneet keskeiset hakusanat. Hakusanat valikoitiin hakupyynnöistä kyselyihin manuaalisesti. Karsinta-algoritmeja sovellettiin sekä tallennus- että hakuvaiheessa. Vertaillut karsintamenetelmät olivat S-algoritmi, muunnettu Lovinsin algoritmi, Porterin algoritmi, taivutuspäätteet poistava Xeroxin algoritmi, sekä taivutuspäätteet ja johtimet poistava Xeroxin algoritmi.

Ensimmäisen vaiheen analyysissä eri vaihtoehtoja verrattiin saannin ja tarkkuuden perusteella. Saanti vakioitiin valituille tasoille, joilla mitattiin keskimääräinen tarkkuus (11 saantitasoa, joista alin oli 0 ja josta edettiin kymme-

nen prosenttiyksikön välein 10 %, 20 % ja niin edelleen aina 100 prosentin saantiin asti). Tämän vertailun tuloksena oli, että kaikki karsintamenetelmät tuottivat suunnilleen 4 - 6 prosenttia paremman tuloksen kuin karsimatta jättäminen. Eri karsintamenetelmien keskinäiset erot kuitenkin jäivät vähäisiksi. Hull (1996) totesi, että tämänkaltaiseen analyysiin ja tuloksiin useimmissa tiedonhakututkimuksissa sitten on tyydyttykin.

Hull (1996) itse jatkoi analyysiä ottamalla huomioon erilaisia tulosjoukon muotoutumiseen vaikuttavia tekijöitä. Yksi hakutulosten analysoinnissa huomioon otettava seikka on tarkasteltavien tulosjoukkojen koko. Tutkimuksissa, joissa tulosjoukkojen dokumentit rankkeerataan eli järjestetään oletetun kiinnostavuuden mukaiseen järjestykseen, valitaan usein jotkin vakiokokoiset tulosjoukot, esimerkiksi 5, 10, 15 ja 30 dokumenttia. Vertailtavien vaihtoehtojen paremmuutta mitataan sitten sen perusteella, millaiset tarkkuusarvot eri vaihtoehdoilla näissä vakiokokoisissa tulosjoukoissa saavutetaan. Mitä korkeampi tarkkuusarvo, sitä parempi tulos. Hull huomautti, että analysoinnissa käytettävien tulosjoukkojen koon pitäisi jotenkin olla suhteessa koko tutkimusaineiston kokoon. Esimerkiksi TREC-hankkeen kaltaisessa valtavassa dokumenttikokoelmassa 10, 50, 100 ja 500 dokumentin joukot olisivat asianmukaisempia kuin edellä esitetyt pienet tulosjoukot. Jos tietokannasta löytyy tietyn kyselyn vastaukseksi satoja relevantteja dokumentteja, kuten TREC-projektin kyselyissä helposti käy, ei eri karsintamenetelmien välille voi löytää minkäänlaista eroa, kun tarkastellaan tulosjoukkoja, joissa on korkeintaan muutamia kymmeniä dokumentteja.

Pienissä tulosjoukoissa vertailtavien vaihtoehtojen erot eivät välttämättä tule näkyviin: kun kysely on tarkkaan rajattu ja hakija käy läpi vain tulosjoukkojen 10 ensimmäistä dokumenttia, saantitaso on alhainen eikä eri karsintamenetelmien ja karsimatta jättämisen välille löydy käytännöllisesti katsoen minkäänlaista eroa. Hullin (1996) mainitsema seikka selittäisi sen, miksi Lennon et al. (1981) ja Harman (1991) eivät havainneet eri karsinta-algoritmien välillä eroja: ensinmainitussa tutkimuksessa tarkasteltiin 5 ja 20 dokumentin kokoisia tulosjoukkoja, jälkimmäisessä 10 ja 30 dokumentin tulosjoukkoja.

Niinpä Hull (1996) tutkimuksensa seuraavassa vaiheessa laati kolme erilaista vertailumenetelmää, joilla suhteutti eri karsintamenetelmiä toisiinsa niiden tarkkuuden tai saannin keskiarvojen perusteella. Nämä vertailut perustuivat vakioituihin saanti- tai tarkkuustasoihin. Eri karsintamenetelmiä ver-

rattiin ja ne rankkeerattiin keskiarvojen mukaiseen järjestykseen. Tämän jälkeen varsinaisessa vertailussa jätettiin huomiotta keskiarvojen väliset absoluuttiset erot ja analyysi perustui vain eri karsintamenetelmien keskinäiseen järjestykseen. Tästä analyysistä saatiin seuraavanlaiset tulokset:

1. Mikä tahansa karsintamenetelmä on hyödyllinen aina, kun kyselyt ovat lyhyitä.
2. Algoritmi, joka karsii johdoksia, on hyödyllinen, kun kyselyt ovat lyhyitä, koska hakusanan kanssa täsmääviä hakemistosanoja löytyy tällöin paremmin; sen sijaan pitkissä kyselyissä johdosalgoritmi tuottaa suhteessa enemmän epärelevantteja dokumentteja.
3. Kun saantitaso on korkea ja kyselyt ovat lyhyitä, Porterin algoritmi toimii hyvin, mutta ei silloin, kun kyselyt ovat pitkiä.

Seuraavaksi Hull analysoi eri karsintavaihtoehtoja kahden tilastollisen merkitsevyystestin avulla, nimittäin varianssianalyysin (ANOVA) ja Friedmanin merkitsevyystestin avulla. Näistä Friedmanin testi osoittautui paremmaksi. Sen avulla voitiin muun muassa osoittaa, että karsimatta jättäminen oli selvästi huonoin vaihtoehto ja että karsinta-algoritmeista S-algoritmi oli selvästi tehottomampi kuin muut. Lovinsin algoritmi oli epätarkin ja tuotti sattumanvaraisia tuloksia: useimmissa kyselyissä se tuotti huonompia tuloksia kuin muut algoritmit, joissain kuitenkin huomattavasti muita parempia. Muiden algoritmien eli Porterin, Xeroxin taivutusmuotoalgoritmin ja Xeroxin johdosalgoritmin välillä ei ollut suuria eroja, kun niitä vertailtiin keskimääräarvojen perusteella.

Yksityiskohtainen kyselyjen ja dokumenttien vertailu paljasti, että samojen pääteainesten karsinta joissakin sanoissa tuotti hyviä tuloksia ja toisissa taas epäonnistuneita tuloksia. Sääntöpohjaiset karsintamenetelmät eivät siis ole paras mahdollinen ratkaisu, vaan karsinnan apuna tarvittaisiin jonkinlaista sanakirjaa. Tosin silloin tullaan riippuvaisiksi sanakirjan kattavuudesta: jos karsittavaa sanaa ei löydy sanakirjasta, algoritmi ei toimi. Esimerkiksi erisnimiä ilmestyy koko ajan uusia. Näitä tuntemattomia sanoja varten olisi siis laadittava oma, sääntöpohjainen algoritminsa. (Hull 1996)

Vielä yksi Hullin (1996) huomio oli, että puhtaan kielitieteellisin periaattein laaditut karsinta-algoritmit eivät aina tuottaneet optimaalisia tuloksia. Tulokset saattaisivat olla parempia, jos karsinta-algoritmeja räätälöitäisiin tiedonhakusovellusten tarpeiden mukaisiksi, mahdollisesti vielä aihealueen erikoistarpeet huomioonottaen.

4.3.4 Muiden kielten karsinta-algoritmit

Edellisissä luvuissa selostetuissa tutkimuksissa kohteena olivat englannin kieltä varten laaditut karsinta-algoritmit. Muiden kielten karsinta-algoritmeilla on saatu parempia tuloksia kuin esimerkiksi Harmanin ja Lennonin et al. vertailuissa. Savoy (1993) kehitti ranskaa varten algoritmin, jonka totesi tuottavan parempia tuloksia kuin karsimatta jättäminen. Savoyn algoritmi tosin oli kehittyneempi kuin Harmanin ja Lennonin et al. vertailemat sääntöpohjaiset algoritmit, sillä siinä käytettiin sanakirjaa katkaisun tulosten tarkistamiseen. Toisessa tutkimuksessaan Savoy (1999) tarkensi, että karsinta hyödytti erityisesti, kun dokumentit olivat lyhyitä. Lisäksi heikko karsinta, jossa käsiteltiin vain taivutusmuotoja, vaikutti hyödyllisemältä kuin myös johtimiin kajoava vahva karsinta; jälkimmäisen ongelmana oli ylikarsinta eli eri sanojen yhdistäminen samaan vartaloon.

Popovic ja Willett (1992) testasivat sloveenin kieltä varten laadittua karsinta-algoritmia ja havaitsivat sen parantavan tiedonhaun tuloksia. Heidän mukaansa pääteainesten karsiminen parantaa tuloksia huomattavasti, mikäli käsiteltävä kieli on morfologisesti mutkikas. Samaan tulokseen tuli myös Kalamboukis (1995), joka tutki nykykreikan päätteiden karsimista. Hänen testiaineistossaan karsinta-algoritmin käyttö paransi sekä saantia että tarkkuutta verrattuna vaihtoehtoon, jossa päätteitä ei karsittu.

Abu-Salemin et al. (1999) tutkimuksen mukaan arabiankielisen tekstin haussa morfologisen normalisoinnin edut ovat selvemmät kuin esimerkiksi englannissa. Myös saksankielisen aineiston käsittelyssä todettiin pääteainesten poistamisen olevan hyödyksi (Kotzias 1990).

Yleensäkin karsinta-algoritmeja ja muita lingvistisiä menetelmiä on 1990-luvulla alettu kehittää enenevässä määrin muitakin kieliä kuin englantia varten. Tämä tietenkin johtuu siitä, että erikielisten tietokoneella tuotettujen ja käsiteltävien tekstien määrä kasvaa valtavasti, jolloin tekstitiedonhaun ongelmiin törmätään kaikilla kielialueilla yhä yleisemmin. Näiden ongelmien ratkomisesta on tullut taloudellisesti kannattavaa muidenkin kielten kuin englannin osalta.

4.4 Leksikko

Leksikaalinen taso voidaan sisällyttää morfologiaan, mutta myös erottaa omaksi kategoriakseen. Saltonin ja McGillin (1983, s. 260) sekä Doszkocsin (1986) mukaan leksikaalisen tason operaatiot kohdistuvat kokonaisiin sanoihin tai sananmuotoihin. Tällaisia ovat esimerkiksi sulkusanalistalla olevien sanojen poistaminen ennen hakemistoon tallentamista, ilmeisten kirjoitusvirheiden korjaaminen (esimerkiksi teh -> the), hakusanan korvaaminen tai sille vaihtoehtoisen ilmauksen lisääminen kyselyyn (esimerkiksi Great Britain, United Kingdom, British Isles, UK, ja Britain ovat haun kannalta synonyymisiä ilmauksia, vaikka niiden merkitys ei täsmällisesti ottaen ole täysin sama) sekä lyhenteiden korvaaminen tai täydentäminen kokonaisella muodolla (YLE -> Yleisradio). Karkeistaen voi sanoa, että leksikaalisella tasolla sanoja ja sananmuotoja käsitellään sanakirjan avulla merkkijonoina, ilman merkitysanalyysiä (joka on paremminkin semantiikan aluetta). Vaikka sanoja ja sanastoja onkin tutkittu monissa tiedonhakututkimuksissa, nämä tutkimukset eivät useinkaan ole perustuneet kielitieteeseen.

Jos kielitieteellisellä leksikkotutkimuksella ei olekaan ollut erityisempää asemaa tiedon tallennuksen ja haun tutkimuksessa, tilanne on muuttumassa. Suurten kieliaineistojen eli tekstikorpusten käyttö on lisääntynyt kirjoitetun kielen tutkimuksessa, kun tekstikorpuksia on yhä enemmän saatavilla tietokoneella luettavassa muodossa. 1980- ja 1990-lukujen vaihteessa korpustutkimuksen suosiota on lisännyt se, että suurtenkin aineistojen käsittely on laite-, ohjelmisto- ja tallennustekniikan (muun muassa CD-tietolevyjen) kehityksen myötä tullut mahdolliseksi mikrotietokoneillakin. Vaikka korpuksia olisi aiemminkin voinut hyödyntää tiedonhakututkimuksessa, se olisi käytännössä ollut liian työlästä, kun aineistoa olisi pitänyt käydä läpi manuaalisesti. (Karlsson 1994).

Korpusten avulla voidaan muun muassa selvittää, esiintyykö jokin ongelmallinen ilmaus ylipäätään teksteissä vai onko se katsottava vain teoreettiseksi tapaukseksi. Ilmaus hyväksytään vain, mikäli sen taajuus korpuksessa ylittää tietyn raja-arvon. Esimerkiksi englannin kielestä on olemassa 200 miljoonaa sanaa käsittävä tekstipankki Birminghamin yliopistossa (Karlsson 1994, s. 266). Tähän Birminghamin korpukseen sisältyy suuri määrä The Times -lehden tekstiä, jota on käytetty muun muassa englannin kielen muuttumisen tutkimiseen. Tätä tarkoitusta varten tietokone poimi nimenomaan raja-arvon alittavat, tuntemattomat sanat tarkempaan analyysiin. Si-

ten löydettiin uudissanoja, jotka eivät sisältyneet perinteisiin sanakirjoihin. (Renouf 1993)

Toinen Birminghamin korpuksen käyttökohde on ollut sanaryhmien (word clusters) erotteleminen tekstistä. Tarkastelemalla jokaisen tekstissä esiintyvän sanan kontekstia voidaan automaattisesti muodostaa sanaryhmiä, jotka kuvastavat tekstin aihepiiriä. Tätä voidaan käyttää hyväksi esimerkiksi tekstin indeksoinnissa ja referoinnissa. Tietokone voi esimerkiksi tuottaa tekstin keskeisistä sanoista ja ilmauksista luettelon, jota indeksoija voi muokata sen sijaan, että hän itse tekstiä lukiessaan laatisi vastaavanlaisen luettelon alusta loppuun manuaalisesti. (Renouf 1993)

Kun suurtenkin korpusten käsittely on tullut aiempaa helpommaksi, tietokonelingvistisiä malleja ja kielitieteellisiä teorioita voidaan testata yhä suuremmilla ja monipuolisemmilla testiaineistoilla. Näin mallien ja teorioiden toimivuus voidaan arvioida entistä luotettavammin. Korpuksia voidaan myös käyttää muuten sääntöpohjaisen analyysin täydentäjänä: jos kieliopin säännöt eivät riitä yksiselitteistämään jotain ilmausta, todennäköisintä tulkintaa voidaan etsiä korpuksen avulla. Jos morfologisessa analyysissä törmätään vaikkapa sananmuotoon öljysäiliöitä, joka voi olla öljysäiliö-sanan taivutusmuoto tai yhdyssana, jonka osat ovat öljysäiliö ja itä, voidaan heuristisesti arvata, että yhdyssanatulkinta on epätodennäköisempi. Päätelmä saa lisävahvistusta, jos korpuksestakaan ei löydy öljysäiliöitä-yhdyssanaa.

4.5 Syntaksi

4.5.1 Yleistä

Syntaksin eli lauseopin tutkimuskohteena on lauseiden rakenne. Lause on kieliopin laajin rakenteellinen yksikkö. Lauseet koostuvat sanoista, jotka ryhmittyvät hierarkkisesti lausekkeiksi. (Karlsson 1994, s. 109)

Syntaktisia ongelmia on tutkittu tiedon tallennuksen ja haun alkuajoista lähtien. Tavoitteena on ollut kehittää tekniikoita, joilla teksteistä voitaisiin automaattisesti tunnistaa sellaisia kahden tai useamman sanan muodostamia ilmauksia (kuten Suomen pankki, hydraulic pump), joilla dokumenttien asiasisältöä tai toisaalta tiedontarvetta voitaisiin kuvata. Ideana on, että tällaisten lausekkeiden tai fraasien avulla voitaisiin kuvata dokumentti tai tiedontarve täsmällisemmin.

Boolen logiikkaan perustuvissa hakujärjestelmissähän hakemistoihin tallennetaan vain yksittäisiä, erillisiä sananmuotoja. Tällöin kadotetaan tieto lauseen syntaktisesta rakenteesta eli sanojen välisistä suhteista. Hakuvaiheessa hakija ei voi ilmaista hakusanojen välisiä suhteita luonnollisella kielellä, vaan hakusanojen suhteet on määriteltävä matemaattis-loogisin keinoin eli Boolen ja läheisyysoperaattoreita käyttäen.

Boolen operaattoreiden ongelmana on, että ne ovat tekstihaussa usein liian epätarkkoja. JA-operaattori ilmaisee vain, että hakusanojen pitää esiintyä samassa dokumentissa. Jotta tiedonhaku pitkistä dokumenteista olisi tarkempaa, monissa hakujärjestelmissä on otettu käyttöön ns. läheisyysoperaattorit. Näiden avulla voi ilmaista esimerkiksi, että hakusanojen on esiinnyttävä samassa virkkeessä tai kappaleessa, tai sitten niiden on sijaittava korkeintaan *N* sanan päässä toisistaan, missä *N* on hakijan määriteltävissä oleva kokonaisluku.

Sanojen esiintyminen lähekkäin ei vielä sekään takaa sitä, että niiden asiasisältö vastaa sitä, mitä hakija oli etsimässä. Suomen vienti Ruotsiin ei ole sama asia kuin Ruotsin vienti Suomeen, vaikka niissä esiintyvät aivan samat sanat samalla etäisyydellä toisistaan - tässä tapauksessa vain Ruotsi- ja Suomi-sanojen sijamuodot ilmaisevat, kumpi on lähtö- ja kumpi kohdemaan roolissa.

Riittävätkö sanojen keskinäisen järjestyksen määrittelevät tai muut vastaavat läheisyysoperaattorit sitten ilmaisemaan hakusanojen väliset suhteet riittävän tarkasti niin, että edellä esitetyn esimerkin kaltaiset lausekkeet voitaisiin erottaa toisistaan, vai tarvitaanko lingvistisiä menetelmiä? Molemmille näkemyksille löytyy kannattajia.

Aikoinaan Salton ja McGill (1983, s. 91) totesivat, että yksinkertaiset syntaktiset menetelmät, jotka perustuivat kontekstivapaaseen lausekerakennekielioppiin (context-free phrase structure grammar), ovat tuottaneet huonompia tuloksia kuin tilastolliset, sanojen esiintymistiheyteen perustuvat menetelmät. Tosin he huomauttavat, että tilanne voi muuttua syntaktisten menetelmien kehittyessä. Myös Sparck Jones (1974) totesi, että kun syntaktisin lauseenjäsennysmenetelmin tuotetut lausekkeet eivät tuottaneet tiedonhaussa sen parempia tuloksia kuin muutkaan tekniikat, tiedonhakututkijat kehittivät niiden sijaan erilaisia tilastomatemaattisia menetelmiä lause-

rakenteen analysoimiseksi, koska nämä vaativat vähemmän tietokone-resursseja.

Myöhemmin Fagan (1989) puolestaan totesi ei-syntaktiset tilastolliset menetelmät riittämättömiksi, kun hän vertasi yksinkertaisia hakusanoja (yhdestä sanasta tai vartalosta muodostuva hakuavain) sisältäneiden kyselyjen tuottamia tulosjoukkoja sellaisiin, joissa hakuavaimena oli useammasta sanasta koostuva sanaliitto tai lauseke. Tällaisia lausekkeita muodostava algoritmi perustui muun muassa sanojen spesifisyyteen ja siihen, miten usein tietyt sanat esiintyivät lähekkäin. Vaikka lausekkeiden käyttö tietyissä tapauksissa selvästi paransi hakujen tulosta, tulokset eivät parantuneet kaikissa testikokoelmissa eivätkä havaitut muutokset yleensä olleet tilastollisesti merkitseviä. Niinpä Fagan piti epätodennäköisenä, että hänen soveltamiaan ei-syntaktisia lausekkeentunnistusmenetelmiä kannattaisi soveltaa tiedonhakujärjestelmissä. Fagan myös analysoi tarkemmin, millaisia lausekkeita tilastolliset menetelmät tuottivat ja totesi, että monien lausekkeiden laatua pystyttäisiin parantamaan melko yksinkertaisillakin syntaktisilla jatkoanalyyseillä. Hänen mukaansa sekä kyselyissä että dokumenteissa esiintyvien sanojen väliset syntaktiset suhteet tulisi ottaa huomioon, kun lausekkeita tunnistetaan. Tässä tutkimuksessaan Fagan ei käytännössä kuitenkaan toteuttanut syntaktista jatkoanalyysiä.

Faganin ja omien tutkimustensa perusteella Croft et al. (1991) toteavat, että lausekkeiden tutkimisessa riittää tiedonhaun tutkimukselle vielä paljon tehtävää. Niiden vaikutusta tiedonhakuun ei riittävästi tunneta. Pitäisikö lausekkeita pitää vastaavanlaisina hakuavaimina kuin yksittäisistä sanoista muodostuvia hakusanoja, vai onko lauseke tulkittava paremminkin hakusanojen väliseksi suhteeksi?

Seuraavissa luvuissa esitellään muutamia tutkimuksia, joissa syntaktisen tason analyysiä on sovellettu lausekerakenteiden tunnistamiseen. Katsaus tutkimuksiin, joissa syntaktisia menetelmiä on kokeiltu erityisesti automaattisessa indeksoinnissa, löytyy Sparck Jonesilta (1974). Muita katsauksia syntaktisista menetelmistä tiedonhakututkimuksessa löytyy muun muassa Schwarzilta (1990) ja Pirkolalta (1999).

Lausekerakenteiden tunnistamisen lisäksi automaattista syntaktista analyysiä voidaan soveltaa esimerkiksi lausetason yksiselitteistämisessä. Siinä pyritään lauseyhteyden perusteella karsimaan pois morfologisessa analyy-

sissä monitulkintaisiksi jääneiden sananmuotojen ylimääräisiä tulkintoja ja päättelemään, mikä tulkintavaihtoehdoista olisi kyseisessä kontekstissa mahdollinen.

Kolmantena tutkimusalueena voidaan mainita hakujärjestelmän ymmärtämän formaalin kyselyn muodostaminen luonnollisella kielellä esitetystä hakupyynnöstä.

4.5.2 Lausekerakenteen automaattisen tunnistaminen

Läheisyysoperaattoreiden avulla voidaan määritellä hakusanojen välinen etäisyys ja usein myös niiden keskinäinen järjestys. Siltikään hakusanojen välisiä suhteita ei aina pystytä määrittelemään riittävän tarkasti ja samalla joustavasti. Jos hakija etsii englanninkielisiä dokumentteja, joiden aiheena on munanvalkuainen, egg white, hänen pitäisi määritellä läheisyysoperaattori niin väljästi, että tulokseksi saadaan myös dokumentit, joissa on ilmaus white of farm eggs. Kyselyn ei kuitenkaan pitäisi tuottaa tulosjoukkoon dokumentteja, joissa on aiheena ovat valkoisten kanojen munat eli white eggs.

Siemensin COPSY-projektissa (Context OPerator SYntax) muodostettiin automaattisen syntaktisen analyysin avulla lausekkeita, joita sitten hyödynnettiin sekä tiedon tallennus- että hakuvaiheessa. Tekstistä tunnistettiin nominilausekkeet ja näistä määriteltiin sanojen väliset dependenssisuhteet[3] eli se, mikä on lausekkeen pääsana ja mikä tai mitkä sanat sen määritteitä. Luvun alussa mainitussa esimerkissä white on pääsana ja egg sen määrite. Nämä dependenssitiedot tallennettiin hakemistoon. Hakuvaiheessa hakupyynnölle tehtiin vastaava dependenssianalyysi. Vastaukseksi kelpuutettiin vain sellaiset dokumentit, joissa sanat esiintyivät samanlaisessa dependenssisuhteessa. White egg hylättäisiin, koska siinä white on egg-sanan määrite eikä päinvastoin. (Schwarz 1990)

COPSY oli osa Siemensin laajempaa TINA-projektia (Text-INhalts-Analyse). TINA-projektin alkuvaiheessa kehitettiin morfologisia menetelmiä, joilla yksittäiset sananmuodot voitaisiin palauttaa normaali- eli perusmuotoonsa. Varsinaisesti sen osaprojekteissa kuitenkin keskityttiin erilaisten

[3] Dependenssin käsitteellä kuvataan sanojen välisiä syntaktisia riippuvuuksia; mikä on pääsana ja mitkä sille alisteisia määritteitä. (Karlsson 1994, s. 143).

syntaktisten ongelmien ratkaisemiseen. Osaprojekteja olivat COPSY:n lisäksi THESYS (THEsaurus SYntax System), jossa tekstistä poimittujen pääsana-määrite-suhteiden perusteella määriteltiin sanoille tesaurussuhteet, ts. määriteltiin laajemmat, suppeammat ja rinnakkaistermit. CLAC-projektissa (CLassification and Abstracting Component) taas järjestelmän tuottamia lausekkeita käytettiin dokumenttien luokituksen ja tiivistelmän teon apuna. REFORM-projektissa (REad it FOR Me) hakujärjestelmä poimi potentiaalisesti kiinnostavasta tekstistä nominilausekkeita ja esitti nämä hakijalle kuvauksena dokumentin asiasisällöstä. Nämä dokumentin sisältöä kuvaavat "asiasanat" siis luotiin vasta hakuistunnon aikana. Niiden perusteella käyttäjän oli tarkoitus päättää, mitkä dokumenteista tulostettaisiin hänelle tarkempaa tutustumista varten. SEMF-projektissa (SEMantic Fields) dependenssirakenteita käytettiin semanttisen verkon rakentamisen apuna. (Schwarz 1990)

Kaikissa edellisissä osaprojekteissa tutkimusaineistona oli englanninkielinen teksti. GEIST-projektissa (German English Information Search Technology) mukana oli myös saksan kieli. GEIST-projektissa dependenssianalyysit tehtiin ensin saksankieliselle hakupyynnölle. Tämän jälkeen kukin normalisoitu lauseke muunnettiin (ei siis varsinaisesti käännetty) joukoksi mahdollisia kohdekielen eli englannin lausekkeita, joilla sitten haettiin englanninkielisestä tietokannasta. (Schwarz 1990) 1990-luvun alussa vastaavia tutkimuksia ei juuri ollut, mutta sittemmin kieltenvälinen tiedonhaku on noussut yhdeksi suosituksi luonnollisen kielen käsittelyn sovellusalueeksi. Tästä osoituksena on muun muassa oman CLIR-ryhmän (cross language information retrieval) perustaminen TREC-hankkeeseen vuonna 1997 (Voorhees & Harman 2000).

Toisena esimerkkinä nominilausekkeita hyödyntävästä tutkimuksesta esitellään Structured Information Management: Processing and Retrieval (SIMPR), jossa kehitettiin morfologis-syntaktisia analyysimenetelmiä puoliautomaattisen indeksoinnin tarpeisiin. Morfo-syntaktisen analyysin perusteella tekstistä poimittiin lausekkeet, jotka muodollisin perustein saattaisivat soveltua sisällönkuvailuun. Tällaiset automaattisesti valikoidut lausekkeet voitaisiin sitten antaa indeksoijan tarkastettavaksi, joka valitsisi niistä kuvailutermeiksi kelpaavat. (Karlsson 1990; vastaava teksti löytyy myös lähteestä Karetnyk et al. 1991)

Esikäsittelyvaiheessa tekstistä muun muassa poistettiin erilaiset typografiset merkinnät sekä pyrittiin tunnistamaan idiomit, jotka tulisi käsitellä kokonaisuuksina eikä sana kerrallaan (kuten in spite of). Sen jälkeen englannin kielen Twol-ohjelma teki morfologisen analyysin, jonka tuloksena saatiin kaikki teoriassa mahdolliset tulkinnat eli luennat (reading). Seuraavaksi ylimääräiset, esimerkiksi homonyymeistä johtuvat, tulkinnat pyrittiin karsimaan pois paikallisen, morfologisen yksiselitteistäminen avulla. Morfologiaan perustuva karsintasääntö saattoi esimerkiksi määritellä, että jos sananmuodolle löytyy useita vaihtoehtoisia yhdyssanatulkintoja, valitaan niistä se, jossa on vähiten yhdyssanarajoja[4]. Karlsson toteaa tämän säännön lähes poikkeuksetta tuottaneen oikean tuloksen (Karlsson 1990).

Seuraavana oli morfosyntaktinen koodaus, jossa lauseen jokaiselle sanalle annettiin syntaktinen koodi. Tämän jälkeen tapahtui varsinainen syntaktinen analyysi ENGCG-formalismin avulla. Rajoituskieliopin (constraint grammar, CG) säännöt perustuvat pitkälti lauseen pintarakenteeseen ja morfologiseen analyysiin. Sen perustana ovat morfologisen analyysin (käytännössä kaksitasomallin) tuottamat luennat. Esimerkiksi englanninkielisessä tekstissä yli 40 prosentille sananmuodoista löydetään morfologisessa analyysissä vähintään kaksi luentaa, usein enemmänkin (Gibb & Smart 1990).

Kun perinteinen lausekielioppi pyrkii löytämään lauseelle yhden ja oikean tulkinnan, rajoituskieliopissa sanoille ja lauseille etsitään ensin kaikki mahdolliset tulkinnat ja näistä karsitaan esiin se oikea. Pintasyntaksin säännöt määrittelevät, miten sananmuoto tietyssä kontekstissa tulkitaan edellyttäen, että määrätyt ehdot täyttyvät. Jos jokin tulkinta ei täytä ehtoja, se hylätään. Samalla periaatteella pyritään löytämään myös lausekkeiden väliset rajat. Viimeisessä vaiheessa kullekin sananmuodolle annettiin (optimitapauksessa vain yksi) tulkinta, jossa sille on annettu syntaktinen, pintatason analyysiin perustuva luokitus. (Karlsson 1990)

Rajoituskielioppi on sääntöpohjainen koodaus- tai disambiguointimenetelmä. Sääntöpohjaiset menetelmät ovat toinen koodausohjelmien päätyy-

---

[4] Edelläkuvatulla tavalla voitaisiin yksiselitteistää oikein esimerkiksi sananmuoto öljysäiliöitä, joka voi olla joko öljysäiliö-yhdyssanan taivutusmuoto tai öljysäiliö- ja itä-sanoista muodostunut yhdyssana. Valituksi tulisi öljysäiliö, koska siinä on vain yksi yhdyssanaraja. Poikkeuksiakin on, esimerkiksi piilevä on monissa teksteissä todennäköisemmin levälaji kuin piillä-sanan taivutusmuoto ja kertasinko voi olla sinkotyyppi eikä kerrata-sanan muoto.

peistä, toinen koulukunta ovat tilastolliset menetelmät. Tilastomenetelmissä käytetään apuna todennäköisyyksiä, jotka lasketaan automaattisesti jo koodatun korpuksen ominaisuuksien perusteella. Tilastollinen koulukunta on pitkään ollut vallitseva, mutta nyttemmin sääntöpohjaiset menetelmät ovat kasvattaneet suosiotaan. Sääntöpohjaiset ohjelmat tyypillisesti tekevät vähemmän virheitä, mutta jättävät osan monitulkintaisuuksista ratkaisematta; tilastolliset menetelmät puolestaan eivät jätä mitään sanaa monitulkintaiseksi, mutta tekevät enemmän virheitä. (Voutilainen 1994; Tolle & Chen 2000).

SIMPR-projektissa ENGCG pystyi karsimaan, tekstin vaikeusasteesta riippuen, ylimääräisistä luennoista pois 95 - 97 prosenttia. Syntaktisen analyysin jälkeen vielä vajaat 10 prosenttia sananmuodoista jäi monitulkintaisiksi eli niihin liittyi enemmän kuin yksi syntaktinen koodi (Karlsson 1990).

Pintasyntaktisen analyysin tulos käsiteltiin edelleen modulilla, joka tunnisti tekstistä sisällön kuvaamiseen mahdollisesti soveltuvia lausekkeita. Lausekkeet normalisoitiin eli eri taivutusmuodot palautettiin haluttuun (hakusana)-muotoon ja sulkusanalistalla olevat sanat poistettiin. Seuraavaksi lausekevarianteista valittiin tarkin vaihtoehto ja jätettiin muut pois; esimerkiksi ilmauksista fuel pump ja pump valittiin ensin mainittu. Lopuksi SIMPR-järjestelmän tuottamat indeksitermit annettiin käyttäjän eli indeksoijan hyväksyttäviksi, korjattaviksi tai hylättäviksi. (Karetnyk et al. 1991; Gibb & Smart 1990)

ENGCG-menetelmää on kehitetty SIMPR-projektin jälkeenkin. Voutilainen (1994) vertasi ENGCG-formalismia muihin koodaus- eli merkintäohjelmiin (tagger), jotka analysoivat englanninkielistä tekstiä ja merkitsevät tekstin rakenneosiin kieliopilliset tunnistekoodit. Voutilaisen mukaan englannin rajoituskielioppi ENGCG oli ainoa sääntöpohjainen koodausohjelma, joka oli kilpailukykyinen verrattuna tilastomenetelmiin perustuviin koodausohjelmiin. ENGCG pystyi yksiselitteistämään eli tuottamaan vain yhden luennan noin 93 - 97 prosentille englanninkielisen tekstin sananmuodoista. Vertailluissa kahdessa tilastopohjaisessa menetelmässä prosenttiluku oli jokseenkin sama, 95 - 97 prosenttia. Toisaalta tilastopohjaiset koodausohjelmat tulkitsivat väärin noin 3 - 4 prosenttia sananmuodoista, kun taas ENGCG tulkitsi väärin vain 0,23 prosenttia sanoista. Se kuitenkin jätti yli kuin 4 prosenttia sananmuodoista monitulkintaisiksi. (Voutilainen 1994)

Seuraavaksi Voutilainen kokeili menetelmiä, joilla voisi yksiselitteistää loputkin ENGCG:n monitulkintaisiksi jättämät sananmuodot. Toisessa näistä menetelmistä laadittiin rajoituskieliopin mukainen kuvaus (tai säännöstö) heuristisista rajoituksista. Toisessa menetelmässä käytettiin jo analysoitua ja yksiselitteistettyä korpuksen osaa jatkoanalyysin apuna. Jos ongelmailmaus täsmäsi johonkin tekstin muussa kohdassa yksiselitteistettyyn ilmaukseen, ongelmailmaus voitiin analysoida tämän muualla jo ratkaistun tiedon perusteella. Molemmilla menetelmillä saatiin monitulkintaisten sanojen osuus pienenemään parin prosenttiyksikön verran ilman, että virheanalyysien määrä olisi juurikaan lisääntynyt.

Voutilainen on sittemmin jatkanut erilaisten analyysiohjelmien kehittämistä. ENGCG:n jälkeen on syntynyt kaupallinen koodausohjelma EngCG-2 ja edelleen EngLite, kevyt englannin kielen jäsennin, jossa analyysin tuloksena saadaan enemmän syntaktista tietoa kuin pelkällä koodausohjelmalla (Conexor 2000). Conexorin funktionaalista dependenssikielioppia (Functional Dependency Grammar of English, FDG), jonka yksi osa EngCG-2 on, on sovellettu myös eräässä TREC-projektissa (Strzalkowski et al. 1998).

Muita lauserakenteen tunnistamiseen ja yksiselitteistämiseen liittyviä tutkimuksia on kuvannut muun muassa Pirkola (1999).

4.5.3 Boolen operaattorien päätteleminen hakupyynnöstä

Yksinkertaisin automaattinen menetelmä muuntaa hakupyyntö kyselyksi on ensin poistaa siitä tarpeettomaksi katsotut sanat sulkusanalistan avulla, karsia jäljellejääneistä sanoista ylimääräiset pääteainekset ja sitten suorittaa haku näillä jäljellejääneillä vartaloilla. Tämä menetelmä on mahdollinen silloin, kun hakujärjestelmä ei edellytä Boolen operaattoreiden käyttöä, vaan kysely muodostetaan jollain muulla periaatteella. Esimerkiksi CITE-näyttöluettelossa kysely laadittiin edelläkuvatulla tavalla siten, että relevanteimmiksi määriteltiin dokumentit, joissa esiintyivät kaikki hakusanat, sitten ne dokumentit, joissa esiintyivät yhtä hakusanaa lukuunottamatta kaikki jne. Alimmalla sijalla olivat dokumentit, joissa esiintyi vain yksi hakusanoista. (Doszkocs 1983)

Jos hakujärjestelmä sen sijaan edellyttää, että hakusanojen väliset suhteet on ilmaistava Boolen operaattoreiden avulla, ei formaalin kyselyn tuottaminen hakupyynnöstä onnistu yhtä yksinkertaisesti kuin CITE-järjestelmässä. Automaattisen menetelmän pitää tällöin osata päätellä, mitkä operaattorit

hakusanojen väliin sijoitetaan. Das-Gupta (1987) tutki, voidaanko luonnollisella kielellä ilmaistuissa hakupyynnössä esiintyvät ja- ja tai-konjunktiot muuntaa automaattisesti sopiviksi Boolen operaattoreiksi. Niiden lisääminen suoraan kyselyyn ei ole mahdollista, koska ne luonnollisessa kielessä eivät ole yksiselitteisiä. Tilanteesta riippuen hakupyynnössä esiintyvä ja voidaan tulkita joko JA-operaattoriksi, kuten tietokoneet ja opetus, tai TAI-operaattoriksi, kuten kissat ja koirat allergian aiheuttajina.

Das-Guptan ideana oli, että jos hakupyynnössä ja-sanalla kytketyt sanat olivat semanttisesti samankaltaisia, niiden välinen kytkentä pitäisi kääntää kyselyyn TAI-operaattoriksi, mutta muussa tapauksessa JA-operaattoriksi. Semanttinen samankaltaisuus pääteltiin vertailemalla hakupyynnön sanojen määritelmiä elektronisesta Webster's English Dictionarystä. Das-Gupta totesi menetelmän toimivan yksinkertaisissa hakupyynnöissä, mutta mutkikkaampiin tapauksiin se ei soveltunut. Menetelmän ongelmana oli myös se, että se oli riippuvainen sanakirjan tasosta. Das-Guptan käytämää sanakirjaahan ei ollut alunperin tuotettu tällaiseen tarkoitukseen, joten sanakirja-artikkeleiden määritelmät eivät olleet niin yhtenäisesti laadittuja kuin automaattisessa vertailussa olisi ollut tarpeen.

4.6 Semantiikka

Semantiikka eli merkitysoppi tutkii sanojen ja niiden yhdistelmien välisiä merkityssuhteita. Tällaisia suhteita ovat muun muassa synonymia, antonymia (vastakkaisuus), anomalia (merkitysten yhteensopimattomuudesta johtuva luonnottomuus) ja moniselitteisyys (Hakulinen & Ojanen 1976).

Synonymian ja homonymian ongelmia on tiedonhaussa perinteisesti pyritty ratkaisemaan asiasanastoilla ja tesauruksilla, jotka eksplisiittisesti määrittelevät indeksitermien väliset suhteet ja ohjaavat niiden käyttöön. Niiden periaatteena on, että jos hakija ei itse valitsemallaan hakusanalla päädy toivottuun tulokseen, hän voi tesauruksesta löytää laajempien, suppeampien ja rinnakkaistermien määrittelyjä, joiden perusteella löytää sopivammat hakusanat tai indeksitermit.

Lingvistinen semantiikka ei tähän mennessä ole tarjonnut kovin monia menetelmiä, jotka olisivat toden teolla pystyneet haastamaan perinteiset tiedonhakututkimuksessa tai alan käytännössä jo vakiintuneet menetelmät. Perinteisestä tesaurusmallista poikkeavia käsitemalleja ja -hierarkioita on

toki kokeiltu. Tutkimushankkeita on ollut monia, mutta tässä esitellään katsauksenomaisesti vain pari: Suhteellisen yksinkertaista tekniikkaa on sovellettu Ranskan puhelinluettelon keltaisten sivujen otsikkohakemistossa (Clemencin 1988). Esimerkkinä mutkikkaammasta sovelluksesta voidaan mainita Tome Searcher (Vickery, 1988; 1989). Yksi mahdollinen vaihtoehto tesauruksille on neuroverkkotekniikka, jonka avulla dokumenteille voidaan tehdä automaattinen sisällönanalyysi ja sen perusteella ryhmitellä samankaltaiset dokumentit yhteen (Honkela et al. 1996; Honkela 1997).

Ranskan MINITEL-puhelinluettelon keltaisten sivujen otsikkohakemistossa oli vuonna 1988 suunnilleen 2 500 erilaista otsikkoa (headings), joita haettiin perinteisesti sanahaulla. Lisäksi otsikoissa esiintyville sanoille oli määritelty rinnakkaisia synonyymejä. Käyttäjät olivat tavallisia tiedontarvitsijoita eli puhelimenomistajia, joille oikeiden otsikoiden löytäminen tuotti vaikeuksia. Liian yleistä hakusanaa käytettäessä otsikoita tuli helposti liikaa ja toisaalta usein ei löytynyt sopivaa otsikkoa lainkaan. (Clemencin 1988)

Edelläkuvattujen ongelmien välttämiseksi rakennettiin hakujärjestelmä, joka perustui tietämyskantoihin. Tässä järjestelmässä haut eivät kohdistuneet suoraan otsikkoihin, vaan sopiva otsikko etsittiin sen kuvauksen (descriptions, indexes) avulla. Ensin käyttäjän syöttämä kysymys analysoitiin suhteellisen karkealla luonnollisen kielen tulkintaohjelmalla. Käyttäjän kysymyksessä esiintyneet sanat ja niiden taivutusmuodot tunnistettiin ja näistä pyrittiin edelleen tunnistamaan myös yhdyssanat. Sen jälkeen käyttäjän syöttämästä lauseesta etsittiin sen pääsana. Jokaista otsikkoa varten tietämyskantaan oli laadittu yksi tai useita päättelysääntöjä, joiden perusteella pyrittiin löytämään oikea tiedontarvetta vastaava otsikko. Näitä 2 500 otsikkoa varten tietämyskannassa oli kaikkiaan noin 20 000 päättelysääntöä eli kuvausta. (Clemencin 1988)

Myös Tome Searcher -järjestelmän käyttäjän oli mahdollista kirjoittaa kyselynsä luonnollisella kielellä. Tämän jälkeen se muunnettiin Boolen logiikan mukaiseksi kyselyksi. Ensin hakupyynnöstä poistettiin sulkulistalla olevat sanat ja lopuista karsittiin pääteainekset pois. Jäljellejääneitä vartaloita etsittiin ohjelman sanakirjasta. Yhdyssanat tunnistettiin, mikäli ne oli määritelty sanakirjaan. Sanakirjan sanat oli määritelty tiettyihin semanttisiin luokkiin. Jos hakupyynnössä oleva sana oli Tome Searcher -ohjelmalle tuntematon, siitä pyydettiin lisätietoja käyttäjältä, kuten tarkistamaan oikein-

kirjoitus, antamaan synonyymi tai valitsemaan semanttinen luokka ohjelman antamasta luettelosta.

Kun Tome Searcher oli käsitellyt jokaisen hakupyynnössä olevan sanan ja muokannut tarpeelliseksi katsotut hakusanoiksi, se muodosti näiden pohjalta formaalin kyselyn. Kyselyä laajennettiin automaattisesti poimimalla välittäjäjärjestelmän sanakirjasta kullekin hakusanalle synonyymit ja vaihtoehtoiset kirjoitusmuodot (Englannin ja Amerikan englannin). Hakusanan rinnalle saatettiin lisätä sen monikkomuoto tai sen loppuun lisättiin katkaisumerkki - englannin kielen vähien taivutusmuotojen vuoksi voitiin tyytyä varsin yksinkertaisiin morfologisiin keinoihin. Yhdyssanat, eli englannin kielen sanaliitot, haettiin yhdistämällä sanaliiton muodostavat sanat toisiinsa läheisyysoperaattorilla. Tome Searcherissa oli myös päättelysääntöjä, joiden perusteella määriteltiin, millä Boolen operaattorilla hakusanat kulloinkin kytkettiin toisiinsa.

Periaatteessa Tome Searcher oli yleinen välittäjäjärjestelmä, jota voisi käyttää missä tahansa tietokannassa. Käytännössä Tome Searcherista oli laadittu toteutus INSPEC-tietokantaa varten. Sanakirja oli laadittu hyödyntämällä INSPEC-tietokannan tesaurusta ja kuhunkin sanakirjan sanaan oli liitetty tieto kyseisen sanan yleisyydestä INSPEC-tietokannassa. Ennen kuin haku toteutettiin, sanakirjasta tarkistettiin, kuinka suuren tulosjoukon kysely tuottaisi. Jos tämä oli suurempi tai pienempi kuin edeltä määritelty rajaarvo, kyselyä supistettiin tai laajennettiin. Tämä tapahtui erilaisin, mukaan lukien semanttisin, menetelmin.

Haun laajentaminen tapahtui esimerkiksi pyytämällä käyttäjää poistamaan kyselystä jokin liian rajaava hakusana tai lisäämällä hakusanaan katkaisumerkki. Semanttinen keino oli lisätä kyselyyn sana, joka ohjelman sanakirjassa oli merkitty hakusanan rinnakkais- tai laajemmaksi termiksi. Teknisesti kyselyä supistettiin kohdistamalla haku vain otsikko- ja asiasanakenttiin, semanttinen supistamiskeino oli korvata hakusana termeillä, jotka sanakirjassa oli määritelty sitä suppeammiksi termeiksi.

Vaikka Tome Searcherin toimintaperiaate olikin lupaava, se ei käytännössä toiminut. Algoritmi, jonka piti verrata hakusanoja tesauruksen termeihin, ei pystynyt löytämään relevantteja termejä. Jotta Tome Searcher olisi toiminut oikean välittäjän tavoin, sen olisi pitänyt kyetä laajentamaan kyselyä synonyymeillä, lyhenteillä, rinnakkaistermeillä ja vaihtoehtoisilla kirjoitusmuo-

doilla. Kun näin ei tapahtunut, tietoa jäi löytymättä liian vähien tai puutteellisten hakusanojen ja hakutermien takia. Semanttisesti suuntautuneen välittäjäjärjestelmän rakentaminen edellyttää, että sen sanakirja on riittävän kontrolloitu ja strukturoitu. Ongelmana on, että tällaisen semanttisen tiedon saattaminen tietokoneella käsiteltävään muotoon on vaikeaa. Tavoite olisikin asetettava alemmas siten, että käyttäjän semanttista tietämystä ei yritettäisi korvata, vaan sen sijaan kehitetään menetelmiä, jotka tukevat käyttäjää hyödyntämään omaa tietämystään. (Sormunen 1989, s. 65)

Jos asiantuntijajärjestelmät eivät osoittautuneetkaan tehokkaiksi tiedonhaun ongelmien ratkaisussa, voidaan semanttiset solmut kenties ratkoa toisilla menetelmillä, joiden lähtökohdat poikkeavat asiantuntijajärjestelmistä, mutta myös perinteisistä tesaurustekniikoista. Esimerkkinä tästä lähestymistavasta on suomalainen WEBSOM-menetelmä, jossa teksti analysoidaan automaattisesti neuroverkkotekniikan, tarkemmin sanoen itseorganisoituvien karttojen avulla (Self-Organizing Maps, SOM). SOM on yleinen itseoppiva järjestelmä tai algoritmi, jolla moniulotteista tilastotietoa voidaan järjestää siten, että tilastollisesti toisilleen läheiset tiedot asettuvat kartassa lähelle toisiaan (Kohonen 1995).

WEBSOM-projektissa testiaineistona olivat englanninkieliset viestit, jotka oli lähetetty neuroverkkoja käsittelevään Usenet-uutisryhmään. Näissä vapaamuotoisissa tekstidokumenteissa esiintyvien sanojen ja niiden kontekstin perusteella laadittiin kaksiulotteinen kartta (map), joka visualisoi tekstikokoelman ja siihen sisältyvien dokumenttien väliset suhteet. (Honkela et al. 1996; Honkela 1997)

Automaattisen sisällönanalyysin ensimmäisessä vaiheessa laadittiin sanakategoriakartta (word category map) dokumenteissa esiintyneiden sanojen ja niiden kontekstin perusteella. Sanat, jotka esiintyivät samanlaisissa konteksteissa, päätyivät kartalla tavallisesti lähelle toisiaan. Seuraavaksi dokumentit koodattiin niin, että kussakin dokumentissa esiintyneet sanat suhteutettiin tähän sanakategoriakarttaan. Eräiden matemaattisten käsittelyjen jälkeen laadittiin SOM-algoritmin avulla koko kokoelmasta dokumenttikartta (document map). Dokumenttikartan vaaleammat alueet ovat osoitus dokumenttikeskittymästä eli siitä, että tuossa kohdassa on useita, hyvin samankaltaisia dokumentteja. Vastaavasti tummat alueet osoittavat, että sillä alu-

eella olevien dokumenttien suhteet toisiin dokumentteihin ovat etäiset eli niiden sisältö ei juuri muistuta toisten dokumenttien sisältöä[5].

Tiedonhaku WEBSOM-dokumenttikartasta muistuttaa enemmän selailua (browsing) kuin perinteistä hakua (search). Kun hakija löytää mielenkiintoisen dokumentin ja haluaa löytää muita samanlaisia, haun laajentaminen ei edellytä uusien hakusanojen tai -termien keksimistä, vaan relevantin dokumentin lähidokumentit yksinkertaisesti poimitaan kartalta lähempään tarkasteluun. Tämäntyyppinen haku on perinteistä menetelmää helpompi esimerkiksi silloin, kun hakija on etsimässä tietoa hänelle oudolta aihealueelta, sillä uusien hakusanojen keksiminen on vaikeaa, jos ala ja sen terminologia ovat outoja. Boolen haulle tyypillistä kaksinaisuutta (dokumentti joko täsmää tai ei täsmää kyselyyn) ei myöskään ole, koska semanttisesti läheisten aihealueiden väliset rajat eivät dokumenttikartalla ole jyrkkiä vaan siirtyminen aihealueesta toiseen tapahtuu pehmeästi. (Honkela et al. 1996)

4.7 Teksti ja diskurssi

Teksti on kielen käytön (ei siis kielen rakenteen) yksikkö. Kielen merkitykset toteutuvat tekstissä - tai tekstinä. Diskurssi tarkoittaa paljolti samaa kuin teksti, mutta erityisesti puheena toteutuvaa vuorovaikutusta. Tekstiin ja diskurssiin liittyviä tutkimusalueita ovat muun muassa tekstin (diskurssin) tilanteinen merkitys, diskurssianalyysi ja tekstin sidoksisuus. (Karlsson 1994)

4.7.1 Tilanteinen merkitys ja diskurssianalyysi

Kielen kontekstista (tilanteesta) riippuvaa käyttöä ja tämän käytön sääntöjä tutkivaa kielitieteen aluetta on perinteisesti nimitetty pragmatiikaksi. Huomioon otettavia seikkoja ovat muun muassa puhujan aikomukset ja uskomukset, kuulijan tieto maailmasta, mitä puhuja olettaa kuulijan tietävän jne. (Hakulinen & Ojanen 1976).

Diskurssianalyysi tarkoittaa lingvistiikassa lähinnä puhekielisen vuorovaikutuksen (interaktion) ominaispiirteiden tutkimusta. Siinä tutkitaan diskurs-

[5] WEBSOM-sovellus on kokeiltavissa osoitteessa <URL: http://websom.hut.fi/websom/ >

sin eri lajeja (keskustelu, väittely, kerronta jne.) ja vuorovaikutusta sääteleviä normeja, esimerkiksi sitä, minkälaisten periaatteiden mukaan keskustelun puheenvuorot vaihtuvat eri keskustelijoiden välillä. Tällaiset normit ovat väljiä eikä niitä pystytä kiteyttämään samanlaisten täsmällisten rakennesääntöjen muotoon kuin esimerkiksi morfosyntaksin sääntöjä. Normit ovat kuitenkin olemassa, niiden kuvaamisessa vain tarvitaan erilaista (sosiologisempaa) otetta kuin perinteisesti morfologian ja syntaksin kuvaamisessa. Diskurssin tutkimus onkin monitieteinen alue, jonka periaatteita on omaksuttu sosiologiasta, sosiaalipsykologiasta, antropologiasta ja erityisesti etnometodologiasta. Siinä kielenkäyttö ymmärretään dynaamiseksi vuorovaikutteiseksi prosessiksi. Arkikieltä analysoidaan sen luonnollisissa yhteyksissä, osana laajempia kulttuurisia käyttäytymiskaavoja. (Karlsson 1994, s. 229 - 230)

Tiedonhakututkimuksessa diskurssianalyysiä on sovellettu muun muassa tiedontarvitsijan ja hänen tiedontarpeensa mallintamisessa. Yksi tällainen malli on ASK (Anomalous State of Knowledge), jota varten analysoitiin tiedontarvitsijan ja välittäjän välistä keskustelua (Brooks & Belkin 1983; Brooks et al 1985, Daniels 1986). Belkinin (1984) mukaan tiedonhankintatilanne on vuorovaikutteista tiedontarvitsijan ja (inhimillisen) välittäjän välistä dialogia. Hänen mukaansa on tärkeää analysoida näiden kahden, erityisesti välittäjän, roolia tässä vuorovaikutustilanteessa ja sen perusteella laatia kognitiivinen malli, jonka avulla voitaisiin muun muassa rakentaa tiedonhakujärjestelmiin tiedonhakua helpottava käyttöliittymä tai välittäjäjärjestelmä.

ASK-mallin kehitystyössä käytiin läpi todellisissa tiedonhakutilanteissa nauhoitettuja keskusteluita ja analysoitiin niistä, miten tiedontarvitsija pyrkii täsmentämään tiedontarvettaan välittäjälle. Tämän perusteella esitettiin joukko kategorioita, joiden avulla tiedontarvitsijan ominaisuuksia voitaisiin kuvata systemaattisesti ja siten säätää tiedonhakua automaattisesti tiedontarvitsijan ominaisuuksien mukaan. Tällaisiksi kategorioiksi todettiin muun muassa se, miten hyvin tiedontarvitsija tuntee aihepiirin, josta tietoa haetaan; tiedontarvitsijan tavoite (esimerkiksi opinnäytetyön kirjoittaminen); työkokemus sekä se, onko tiedontarvitsijalla aiempaa kokemusta suorakäyttöjärjestelmistä tehdyistä tiedonhauista. (Daniels 1986)

Diskurssianalyysiä on käytetty myös informaatiotutkimuksen itseymmärryksen tutkimiseen. Green (1991) poimi LISA-tietokannan referaateista virk-

keitä, joissa esiintyi sana information ja analysoi, miten ne heijastavat informaatiotutkijoiden kognitiivisia malleja. Greenin päätelmänä oli, että informaatiotutkijoiden kognitiivinen malli tiedonhausta on mekanistinen ja eikä tiedontarvitsijoiden kognitiivisia malleja taas oteta riittävästi huomioon. Hänen mielestään olisi siirryttävä järjestelmäpainotteisesta näkökulmasta käyttäjäpainotteiseen näkökulmaan ja tutkittava enemmänkin sitä, miten tietojärjestelmissä oleva tieto todella siirtyy käyttäjilleen eli tiedontarvitsijoille. Tähän suuntaan informaatiotutkimuksessa onkin 1990-luvulla jo siirrytty (ks. Froehlich 1994).

4.7.2 Tekstin sidoksisuus

Tyypillinen teksti on rakenteellisesti sidoksinen kokonaisuus. Sidoksiseen tekstiin kuuluvien lauseiden välillä on erilainen viittaussuhteiden verkko. Tekstin sidoksisuus eli koheesio ilmenee yleensä leksikaalisten sidoskeinojen käyttönä, ennen kaikkea sanojen valintana, sekä vapaiden morfeemien että sopivien pronominien käyttönä. (Hakulinen & Ojanen 1976; Karlsson 1994)

Tekstin sisäiset viittaukset voivat tapahtua esimerkiksi anaforien, kataforien ja ellipsien avulla. **Anaforinen** ilmaus viittaa taaksepäin, tekstissä edellä mainittuun seikkaan, esimerkiksi Pekka meni kauppaan. *Hän* osti maitoa. Tyypillisiä esimerkkejä anaforista ovat pronominit. **Katafora** vastaavasti viittaa tekstissä jäljempänä seuraavaan seikkaan. **Ellipsissä** identtinen aines jätetään sen toistuessa pois, esimerkiksi Matti juoksee nopeammin kuin Pekka ∅ (juoksee-sanaa ei toisteta, vaan lause on siltä osin tyhjä, ∅). (Hakulinen & Ojanen 1976; Karlsson 1994) Anaforista, kataforista, ellipseistä ja muista vastaavista ilmauksista käytetään seuraavassa yksinkertaisuuden vuoksi anafora-yhteisnimitystä.

Anaforia käytetään paljon sekä kirjoitetussa että puhutussa tekstissä. Niiden tarkoituksena tekstin elävöittäminen siten, että samaa ilmausta ei toisteta sellaisenaan jatkuvasti, vaan alkuperäistä ilmausta muuntaen. Anaforien avulla tekstissä muualla mainittua seikkaa ei tarvitse toistaa. Toisaalta anaforat parantavat tekstin sidoksisuutta, kun tekstissä viitataan aiemmin puheena olleeseen seikkaan. Anaforan tulkinnassa eli avaamisessa (resoluutiossa) kuulijan pitää ratkaista tekstin viittaussuhteet, eli päätellä **korrelaatti**, johon anafora viittaa. Anaforan avaamisessa voidaan tarvita niin

morfologisen, syntaktisen, semanttisen kuin pragmaattisenkin tason tietoja. (Liddy 1990, Pirkola 1996)

Liddy et al. (1987) olivat ensimmäisiä, jotka tutkivat anaforan vaikutusta tiedon tallennukseen ja hakuun. He toteavat, että vaikka ihmiset yleensä pystyvät avaamaan anaforat suuremmitta vaikeuksitta, niiden avaaminen tuottaa ongelmia alueilla, joissa tekstin sisältämät viittaussuhteet olisi pääteltävä automaattisesti tietokoneohjelmalla. Tällaisia alueita ovat: luonnollista kieltä ymmärtävät järjestelmät (Natural Language Understanding, NLU), joissa tietojärjestelmä luo käsiteltävänä olevasta tekstistä semanttisen mallin tai esitystavan; kysymys-vastaus-järjestelmät (Question Answering); tiivistelmän teko automaattisesti (Automatic Extracting); tiedontarvitsijan esittämän hakupyynnön analysointi (Query Analysis) ja sanojen esiintymistaajuuteen perustuva tiedonhaku.

Liddy ja kumppanit tutkivat näistä aihealueista viimeistä eli anaforien vaikutusta tiedonhakuun. Ensiksi laadittiin luokittelu siitä, minkätyyppiset englannin kielen ilmaukset voisivat olla tulkittavissa anaforiksi. Tämän perusteella laskettiin anaforien yleisyys kahden kirjallisuusviitetietokannan referaateissa sekä arvioitiin, millaisilla algoritmeilla näistä löydettyjen anaforien avaaminen voitaisiin suorittaa automaattisesti. (Liddy et al. 1987; Liddy 1990). Kun anaforien vaikutusta tiedonhakuun analysoitiin, tulokset olivat ristiriitaisia. Joissain haku- ja anaforatyypeissä anaforan avaaminen paransi hakutuloksia, toisissa tapauksissa taas huononsi, joissain taas ei vaikuttanut mitenkään. Tämän vuoksi Liddy (1990) ei pystynyt esittämään periaatteita sille, miten anaforia tulisi tiedonhakujärjestelmissä käsitellä.

Pirkolan tutkimuksessa Liddyn ja kumppanien luetteloon lisättiin vielä yksi mahdollinen ellipsien ja anaforien avaamisen sovellusalue, nimittäin läheisyysoperaatioihin perustuva tekstihaku. Kun tiedonhaun alaa rajataan läheisyysoperaatioilla niin, että hakuavainten on esiinnyttävä samassa virkkeessä tai samassa kappaleessa, ei kysely tuota toivottua tulosta, jos vaikkapa kahdesta hakuavaimesta toinen on korvattu ellipsillä tai anaforalla. (Pirkola 1996, Pirkola & Järvelin 1996a)

Pirkola tutki anaforien vaikutusta läheisyysoperaatioita sisältäviin kyselyihin siten, että ensin haettiin hakuavaimet yhdistettynä JA-operaattorilla. Tämän jälkeen tästä tulosjoukosta poistettiin sellaiset dokumentit, jotka saatiin yhdistämällä hakuavaimet joko virke- tai kappaleoperaattorilla. Jäljelle-

jääneestä tulosjoukosta tutkittiin, olivatko virke- tai kappaleoperaattorilla löytymättä jääneet dokumentit sellaisia, että niissä varsinaisen hakuavaimen sijasta oltiinkin käytetty sen anaforaa tai ellipsiä.

Anaforien ja ellipsien vaikutusta analysoitiin yksityiskohtaisesti erilaisissa kahden hakuavaimen muodostamissa hakutyypeissä. Hakuavaintyypit olivat seuraavat: yksittäinen sana erisnimenä ja yleisnimenä, yhdyssana erisnimenä ja yleisnimenä, henkilönnimi sekä sanaliitto erisnimenä ja yleisnimenä. Tutkimuksen tuloksena todettiin, että anaforien ja ellipsien resoluutio parantaa hakutuloksia, kun haetaan henkilönnimiä, ja ellipsien resoluutio myös silloin, kun ainakin yksi hakuavain on tyypiltään erisniminen sanaliitto. Näissä tapauksissa saanti oli ellipsien ja anaforien käytön vuoksi alempi kuin tilanteessa, jossa ellipsien ja anaforien tilalla olisi ollut varsinainen hakuavain. Erisnimiellipsien resoluutio paransi saantia virkeoperaattoria käytettäessä noin 38 % ja kappaleoperaattoria käytettäessä noin 29 %. Kun taas vastaavien erisnimien anaforat avattiin, saanti oli virkeoperaattoria käytettäessä noin 18 % ja kappaleoperaattoria käytettäessä noin 11 % parempi kuin ilman avausta. Muissa hakuavaintyypeissä resoluution vaikutus oli vähäinen. (Pirkola 1996, Pirkola & Järvelin 1996a)

Kiinnostava tulos oli, että sekä saanti että tarkkuus parantuivat, kun sukunimiellipsit ja henkilönnimien anaforat avattiin - yleensähän saanti ja tarkkuus ovat toisilleen käänteisiä ilmiöitä. Erisnimisten sanaliittojen ellipsien resoluutio saattoi joissain tapauksissa laskea tarkkuutta, mutta yleensä tarkkuus kuitenkin parani; koska resoluutio kuitenkin paransi saantia, resoluutio kaiken kaikkiaan oli hyödyllistä. (Pirkola 1996; Pirkola & Järvelin 1996b) Tutkimustulos on tärkeä siinä mielessä, että jos vain tietyntyyppisten anaforien ja ellipsien avaamisella on vaikutusta tiedonhaun tuloksiin, kannattaa automaattisia avaamismenetelmiä kehittää vain näitä erikoistapauksia varten. Tämä on taloudellisesti halvempaa ja todennäköisesti myös tuottaa täsmällisempiä tuloksia.

5 TUTKIMUKSESSA KÄYTETYT OHJELMISTOT

Luvussa 1 rajattiin yleisellä tasolla tutkimuksen kohteeksi käänteishakemistoon ja Boolen logiikkaan perustuva tiedonhakujärjestelmä. Ennen kuin tutkimushypoteesit tarkennetaan seuraavassa luvussa 6, esitellään ensin tässä luvussa se konkreettinen ohjelmistojoukko, jota tutkimuksessa käytettiin: tiedonhakujärjestelmä ja suomen kielen morfologiset tulkintaohjelmat.

5.1 BASIS-hakujärjestelmä

BASIS on Open Text Corporationin markkinoima tiedonhakujärjestelmä, joka on saatavilla useisiin eri tietokoneisiin ja käyttöjärjestelmiin (esimerkiksi Windows NT, IBM AIX ja Sun Solaris)[1]. BASIS tuli markkinoille vuonna 1968 ja on laajalti eri puolella maailmaa käytetty ohjelmisto. BASIS on perinteinen ja vakiintunut tiedonhakujärjestelmistä, johon on vuosien varrella kehitetty varsin monipuoliset hakumahdollisuudet. Kuten kaikissa suurissa kaupallisissa hakujärjestelmissä, tiedonhaku perustuu käänteistiedostoihin ja Boolen logiikkaan. Tässä mielessä BASIS-järjestelmää voi pitää tyypillisenä tiedonhakujärjestelmänä ja siinä mielessä sopivana alustana FULLTEXT-projektin tyyppiseen tutkimukseen.

Tässä tutkimuksessa käytettiin BASIS Text Information Management -ohjelmiston K-versiota. K-versio oli toteutettu modulaarisesti, jolloin siihen voitiin suhteellisen helposti liittää muita ohjelmia, kuten suomen kielen tulkintaohjelmia. Hakujärjestelmästä tuli vuonna 1990 markkinoille uusi versio, BASISplus (tai BASIS-L), johon oli yhdistetty sekä teksti- että relaatiotietokantojen ominaisuudet. FULLTEXT-projektin aikana suomen kielen morfologisia tulkintaohjelmia ei voinut liittää BASISplus-ohjelmistoon, joten projektissa käytettiin BASIS-K-versiota.

BASIS-K koostui keskusjärjestelmästä ja useista irrallisista moduleista, joita voitiin liittää tiedonhakujärjestelmään tarpeen mukaan. Keskusjärjestelmä (central system) oli perusosa, jolla muodostettiin hakemisto sekä haet-

---

[1] Tämän väitöskirjan julkaisemisen aikana BASIS-ohjelmiston tuorein versio oli 8. Aikoinaan BASIS-järjestelmän tuottaja oli Information Dimensions, Inc. Sittemmin Open Text Corporation osti yrityksen, ja Information Dimensions on nyt Open Text BASIS Division. BASIS-ohjelmiston versiossa 8 on muun muassa erilaisia World Wide Webin käyttöä tukevia ominaisuuksia. <URL: http//:www.InformationDimensions.com/> ja <URL: http://www.opentext.com/basis/>

tiin ja selattiin tietoa, kun taas modulit olivat lisätoimintoja varten. Muun muassa tietojen syöttöä ja muokkausta sekä tesaurusten laatimista ja selausta varten oli omat modulinsa. (Salminen 1988; Davies 1991) BASIS-K tallensi sille syötetyn tekstin dokumentteina, jotka puolestaan jaettiin kenttiin tietokannan tuottajan määrittelemällä tavalla. Kenttiä saattoi olla vähintään yksi ja korkeintaan 2 000 (Bain et al. 1989).

BASIS-K-ohjelmassa oli tarjolla useita tapoja poimia kentästä halutut tiedot hakemistoon. Yksi mahdollisuus oli ottaa kentän sisältö hakemistoon kokonaisuudessaan, yhdeksi hakemiston merkkijonoksi. Tällä tavalla yleensä tallennettiin kirjoittajan nimet ja päivämäärät. Tällöin esimerkiksi kirjoittajan etu- ja sukunimen välissä oleva välilyönti otetaan myös huomioon hakemistossa. Toisaalta vapaamuotoinen teksti - kuten esimerkiksi pitkät lehtiartikkelit - tallennettiin yleensä siten, että tekstiä sisältävästä kentästä poimittiin jokainen sananmuoto erikseen ja nämä yksittäiset sananmuodot tallennettiin hakemistoon (free-text indexing). Tässä tapauksessa sanojen välillä olevia välilyöntejä ei tallennettu hakemistoon kuten edellisessä vaihtoehdossa. BASIS-ohjelmistossa oli lisäksi mahdollista määritellä sulkusanalista (stop word list), jolla lueteltuja sanoja ei otettu mukaan hakemistoon. Oli myös mahdollista, että kentän sisältöä ei tallennettu hakemistoon ollenkaan. (Bain et al. 1989; Salminen 1992)

Hakemistoon tallennettuun sananmuotoon liitettiin osoite, joka ilmaisi tarkemmin, missä dokumentissa ja missä osassa dokumenttia kyseinen sananmuoto oli esiintynyt. BASIS-K-versiossa tarkin mahdollinen osoite oli virke. Virkkeen rajaksi tulkittiin piste, huutomerkki tai kysymysmerkki, jota seurasi välilyöntimerkki. BASISplus-versiossa osoitteet voitiin ilmaista vielä tarkemmin eli yksittäisen sanan tarkkuudella. Käyttäjän oli mahdollista selata hakemistoja. (Saffady 1989; Bain et al. 1989; Salminen 1992)

BASIS-K-hakujärjestelmässä hakusanoina voitiin käyttää täsmällisiä merkkijonovakioita (esimerkiksi *lomaosake*), mutta hakusanat oli myös mahdollista katkaista oikealta (*lomaosak**) tai vasemmalta (**osake*) merkkijonokaavioiksi, jolloin voitiin hakea tietyllä tavalla alkavia tai päättyviä sanoja. Jokeri- ja peittomerkkejä oli tarjolla useita erilaisia. Jokerimerkeistä käytetyin lienee *-merkki, joka korvasi hakusanassa rajoittamattoman määrän muita merkkejä. (Saffady 1989)

Hakusanat voitiin kytkeä toisiinsa Boolen operaattoreilla (JA, TAI, EI) tai läheisyysoperaattoreilla. Läheisyysoperaattorien käytöstä on tietenkin hyötyä vain silloin, kun hakemistosanojen osoitteet on tallennettu tarkemmin kuin koko dokumentin tasolla, eli osoite ilmaisee myös sanan tarkemman sijainnin dokumentin sisällä. BASIS-K-versiossa tietokannan tuottaja voi määritellä, tallennettiinko osoitteet kappaleen vai virkkeen tarkkuudella - molempia tallennustapoja ei voinut käyttää samanaikaisesti (Salminen 1992). Hakija ei siis voinut vapaasti määritellä, hakeeko hän samassa virkkeessä vai samassa kappaleessa esiintyviä sanoja, vaan hänen oli mukauduttava tietokannan tuottajan ratkaisuun.

Numeerisia tietoja (kuten vuosiluvut) voitiin tallentaa erikseen numeerisiin kenttiin ja niiden hakemisessa voitiin käyttää hyväksi matemaattisia operaattoreita (=, >, <). Käyttäjä voi halutessaan kohdistaa kyselyn vain tiettyihin kenttiin. (Saffady 1989; Bain et al. 1989)

BASIS-K-hakujärjestelmässä oli mahdollista hakea myös suoraan tekstitiedostosta eli tehdä peräkkäishaku. Tällöin kysely voitiin kuitenkin kohdistaa vain valmiiseen tulosjoukkoon, joka oli ensin rajattu tekemällä normaali kysely hakemistoon. Koko tietokantaan peräkkäishakua ei voinut kohdistaa. Hakusanojen täsmällinen keskinäinen järjestys voitiin määritellä vain peräkkäishaussa (esimerkiksi, että hakusanojen *kansan* ja *uutiset* pitää esiintyä nimenomaan tässä järjestyksessä). Hakemistossa osoitteet pystyttiin tallentamaan korkeintaan virkkeen tarkkuudella eikä sananmuotojen keskinäistä sijaintia virkkeen sisällä siten tiedetty. Se voitiin selvittää vain tutkimalla itse dokumentin teksti, jossa sananmuotojen alkuperäinen järjestys oli tallella. (Salminen 1988, 1992)

Hakutilanteessa BASIS tutki hakemiston ja ilmoitti sitten osumatiedot, ts. kuinka monessa tietokannan dokumentissa kukin hakuavain oli esiintynyt, sekä muodosti näistä dokumenteista tulosjoukon (kori, setti). Kullakin tulosjoukolla oli oma yksilöity tunnuksensa, jonka avulla tulosjoukkoja voitiin valita ja yhdistellä haun myöhemmissä vaiheissa. Käyttäjä voi tulostaa tulosjoukon joko kokonaisuudessaan tai yksilöidä ensin halutut tietueet tulosjoukon ja tietueen tunnusten avulla. (Saffady 1989; Bain et al. 1989)

FULLTEXT-projektissa BASIS-ohjelmistosta oli käytössä versio BASIS-K R520.33. Testilaitteistona oli VAX 11/750, jonka käyttöjärjestelmänä oli VAX/VMS:n versio V4.2. BASIS-K:n suomalaiseen versioon oli jo vuonna

1987 liitetty Pascal-ohjelmointikielellä toteutettu Morfo-ohjelma, jota oli hyödynnetty useiden sanomalehtiarkistojen tuottamisessa. FULLTEXT-projektissa Morfosta kuitenkin käytettiin testaushetkellä tuoreinta versiota, joka oli C-kielinen. Myös muista kolmesta tulkintaohjelmasta käytettiin tuoreinta saatavilla ollutta versiota.

5.2 Suomen kielen morfologiset tulkintaohjelmat

Suomen kielen automaattista analyysiä on tutkittu kahdella taholla: SITRAn Kielikone-projektissa sekä Helsingin yliopistossa yleisen kielitieteen laitoksella ja tietokonelingvistiikan tutkimusyksikössä. Kummankin tahon tutkimus- ja kehitystoiminnassa syntyneiden ohjelmien tuotteistamista ja markkinoimista varten on perustettu omat kaupalliset yritykset: Kielikone-projektin ohjelmistoja markkinoi Kielikone Oy ja yleisen kielitieteen laitoksen ohjelmia Lingsoft Oy.

Molemmat tutkimusryhmät ovat tuottaneet joukon erilaisia suomen kielen tulkintaohjelmia, joita on käytetty muun muassa tekstin automaattiseen tavuttamiseen ja oikeinkirjoituksen tarkistamiseen. FULLTEXT-projektissa hyödynnettiin vain sananmuotoja käsitteleviä eli morfologisia ohjelmia, joista niistäkin pelkästään kahta tyyppiä:

- perusmuodoista taivutusvartaloita tuottavat ohjelmat (Hahmotin, Finstems)
- taivutusmuotoja perusmuotoon palauttavat ohjelmat (Morfo, Twol)

Seuraavissa luvuissa esitellään ohjelmat lyhyesti sekä annetaan esimerkki ohjelmalle annettavasta syötteestä ja sen tuottamasta tulosteesta. Jokaisen ohjelman tuottamia tulosteita on kuitenkin mahdollista hienosäätää toisenlaiseksi - esitystapa riippuu siitä, mitä tietoja ja missä muodossa kulloisessakin tietojärjestelmässä tarvitaan. Esimerkiksi tässä tutkimuksessa usein mainittu käsite **perusmuoto** ei sinänsä ole itsestäänselvyys, vaan määrittelykysymys. Kaikissa näissä neljässä morfologisessa ohjelmassa perusmuoto vastasi suomen kielen sanakirjoissa yleisesti käytettävää **sanakirjamuotoa** eli hakumuotoa: nomineilla (substantiivit, adjektiivit, pronominit, lukusanat) se on yksikön nominatiivi (kuten kissa, pulma, se, kaksi, vihreä) ja verbeillä I infinitiivi (istua, tulla) (Karlsson 1994, s, 171).

5.2.1 Taivutusvartaloita tuottavat ohjelmat

Taivutusvartaloita tuottavien ohjelmien syötteenä on sana perusmuodossaan. Tästä ohjelma tuottaa eri taivutusvartalot eli kaikki ne mahdolliset vartalot, joissa sana eri taivutusmuodoissaan voi esiintyä (esimerkiksi kauppa -> kauppa-, kaupa-, kauppo-, kaupo-).

FINSTEMS

Finstems on Helsingin yliopistossa kehitetty suomenkielisten substantiivien, adjektiivien ja verbien taivutusvartaloita tuottava ohjelma. Sen aiempi, vain substantiiveja käsittelevä versio oli liitetty useisiin tiedonhakujärjestelmiin, muun muassa IBM:n STAIRS-hakujärjestelmän suomalaiseen versioon.

Syöte: Perusmuodossa oleva sana, jonka edessä on sanaluokan tunnus (N:substantiivi, A:adjektiivi, V:verbi), esimerkiksi N:rauha. Lisäksi käyttäjän on yhdyssanoissa merkittävä viimeisen sanan eteen kauttaviiva, esimerkiksi maailman/sota.

Tuloste: Syötetyn sanan taivutusvartalot, kuten rauha, rauhoi, rauhoj.

Finstems päättelee sanan taivutusluokan sen kirjoitusasun, lähinnä sanan viimeisten kirjainten perusteella, joten se ei tarvitse toimiakseen varsinaista sanakirjaa. Tosin eräät suomen kielen sanat taipuvat poikkeavasti, joten analyysin aluksi Finstems tarkistaa poikkeavien sanojen luettelostaan, kuuluuko käyttäjän antama sana tähän ryhmään. Muutoin Finstems tuottaa vartalot muokkaamalla perusmuotoa säännöillä, jotka ottavat huomioon taivutusluokan, vartalon tyypin (esim. yksikkö vai monikko) ja suomen kielen yleiset astevaihtelusäännöt.

Finstems käsittelee syötetyt sanat vasemmalta oikealle. Yhdyssanojen oikea käsittely edellyttää, että käyttäjä on merkinnyt, mistä yhdyssanan viimeinen osa alkaa. Tämä siksi, että yhdyssanoissa yleensä vain sanan loppuosa taipuu. Toisaalta sanan taipumiseen vaikuttaa myös sen tavujen lukumäärä. Koska Finstems ei tunnista itse sanoja, se voi päätyä väärään taivutusluokkaan, mikäli laskee yhdyssanan kaikki eikä vain viimeisen osan tavut.

Eräissä tapauksissa sanan ulkoasu ei ole tarpeeksi yksiselitteinen täsmällisen taivutusluokan määrittelemiseen. Tällaisia ovat monet s-päätteiset sanat (esimerkiksi tilus, pakkaus). Tällöin Finstems tuottaa varmuuden vuoksi

kaikki mahdolliset taivutusvartalot, jotka taivutusluokkaan kuuluvat. Joissain tapauksissa näin tulee tuotettua ylimääräisiäkin vartaloita, joita todellisuudessa ei esiinny, kuten rakkaus -> rakkaude, mutta pakkaus -> pakkaude. (Koskenniemi 1985b)

HAHMOTIN

Hahmotin on SITRAssa kehitetty ohjelma, joka tuottaa taivutusvartalot suomen kielen substantiiveille, adjektiiveille ja verbeille.

Syöte: Perusmuodossa oleva sana. Oletuksena on, että sana on substantiivi, muutoin käyttäjän on ilmoitettava sanaluokka (a adjektiivi, v verbi), esimerkiksi a tarkka.

Tuloste: Syötetyn sanan taivutusvartalot, kuten rauha, rauho.

Hahmotin-ohjelman toimintaperiaate on pääperiaatteissaan sama kuin Finstems-ohjelmankin. Kun taivutusluokka ei ole täysin varma, myös Hahmotin tuottaa kaikki periaatteessa mahdolliset taivutusvartalot, vaikka näin tuotettaisiinkin myös ylimääräisiä muotoja. Hahmotin käsittelee syötetyn sanan oikealta vasemmalle, minkä ansiosta yhdyssanojen osia ei tarvitse merkitä millään tavalla. Yhdyssana-analyysi tuottaa oikean tuloksen, vaikka yhdyssanat syötetään ohjelmalle samalla tavalla kuin yksiosaiset perussanat.

5.2.2 Perusmuoto-ohjelmat

Perusmuoto-ohjelmat sisältävät sanaston ja suomen kielen taivutussäännöt, joiden perusteella ne pystyvät päättelemään, mikä on jonkin sananmuodon perusmuoto (esimerkiksi kodeissamme -> koti). Lisäksi ne kykenevät jakamaan yhdyssanat osiinsa (kuten ryhmähenkivakuutusta -> ryhmä + henki + vakuutus). Käyttötarpeen mukaan ohjelmat voivat perusmuodon lisäksi antaa muitakin morfologisen analyysin tuloksia, kuten taipuneen sananmuodon sanaluokan ja sijamuodon. Näitä tuloksia voidaan kutsua termillä **luenta** (reading).

Suomen kielen taajuussanaston yleisimmät sanat (eli näiden eri sananmuodot yhdessä) kattoivat valtaosan taajuussanaston kokoamiseen käytetyistä teksteistä. 100 yleisintä sanaa kattoi teksteistä 35,1 prosenttia, 1 000 yleisintä sanaa 64,8 prosenttia ja 10 000 yleisintä sanaa 89,4 prosenttia teksteistä (Saukkonen et al. 1979, s. 21). Kun kerran tekstit sisältävät enimmäkseen

yleisimpien sanojen eri sananmuotoja, voidaan arvioida, että pelkästään yleisimmät sanat sisältävä perusmuoto-ohjelman sanakirja riittää varsin kattavaan analyysiin.

Mitä suurempi perusmuotoihin palauttavan ohjelman sanakirja on, sitä suurempi osuus sananmuodoista tunnistetaan, mikä tietenkin on eduksi. Haittana kuitenkin on, että samalla yhä useammalle sanalle löydetään useampi kuin yksi tulkinta (luenta). Esimerkiksi ilmaisin voi olla joko substantiivin nominatiivimuoto (vrt. genetiivi ilmaisimen), verbin yksikön ensimmäisen persoonan imperfekti (vrt. I infinitiivi ilmaista) tai adjektiivin superlatiivi (vrt. positiivi ilmainen); albumiini puolestaan voi olla substantiivin nominatiivimuoto ('eräs valkuaisainelaji', vrt. genetiivi albumiinin), tai substantiivin illatiivimuoto ensimmäisen persoonan omistusliitteineen (vrt. nominatiivi albumi).

Käytännössä sanakirjaa ei siis kannata kasvattaa miten suureksi tahansa, jottei ylitulkintojen määrä kasvaisi liikaa. On pyrittävä löytämään optimikoko, jolla sekä tunnistamattomien että monitulkintaisten sananmuotojen määrä jää mahdollisimman pieneksi.

MORFO

Morfo on SITRAssa kehitetty ohjelma, jonka sanakirjassa oli FULLTEXT-projektin tutkimushetkellä noin 60 000 perussanaa. Näiden lisäksi se pystyy tunnistamaan näitä perussanoja sisältävät yhdyssanat. Analyysi etenee sananmuodon lopusta alkuun eli oikealta vasemmalle.

Syöte: Yksittäinen sananmuoto, esimerkiksi rajoituksen, tai tekstitiedosto.

Tuloste: Syötetyn sanan perusmuoto, kuten rajoitus (haluttaessa myös koko luenta, joka sisältää sanaluokka-, sijamuoto- yms. tiedot, kuten rajoituksen-sananmuodosta: Noun SG Gen).

Morfo tutkii analysoitavaa sananmuotoa suomen kielen taivutussääntöjen ja sanakirjansa (sanastonsa) pohjalta. Aluksi Morfo tarkistaa ns. välisanakirjasta, onko sananmuoto jonkin taipumattoman tai vähän taipuvan sanan esiintymä. Mikäli näin ei ole, Morfo seuraavaksi jaottelee sananmuodosta kaikki mahdolliset pääteainekset. Näin saatuja sananvartaloita etsitään Morfon toisesta välisanakirjasta. Mikäli vartaloa ei tunnisteta, muokataan vartaloita edelleen tutkimalla muun muassa, onko käsiteltävässä sanassa aste-

vaihtelua. Muokkauksella saatuja sananvartaloita verrataan sitten Morfon varsinaiseen sanakirjaan. Kun sieltä löydetään muokkauksella tuotetun sananvartalon kanssa täsmäävä vartalo tai vartalot, Morfo tulostaa näitä vartaloita vastaavan perusmuodon analyysinsä lopputulokseksi. (Jäppinen et al. 1983)

Mikäli analysoitavalle sananmuodolle ei löydy vastinetta Morfon sanakirjasta, Morfo kokeilee vielä yhdyssana-analyysiä. Tällöin Morfo lähtee sananmuodon lopusta etsimään merkkijonoa, jonka voi tulkita jonkin sanakirjassa olevan sanan esiintymäksi. Mikäli tällainen merkkijono löytyy, Morfo palauttaa vastaavan sanan perusmuodon ja siirtyy tutkimaan sananmuodon alkuosaa omana itsenäisenä merkkijononaan. Jos sananmuotoa ei pystytä tunnistamaan yhdyssana-analyysinkään avulla, se joko puuttuu Morfon sanakirjasta tai siinä oleva kirjoitusvirhe estää analyysin onnistumisen. (Jäppinen et al. 1983)

TWOL

Twol (tai Fintwol) on Helsingin yliopistossa kehitetty, kaksitasomalliin perustuva ohjelma, joka pystyy sekä tunnistamaan että tuottamaan suomenkielisten sanojen taivutusmuotoja. FULLTEXT-projektissa ei testattu Twol-ohjelman ominaisuuksia taivutusmuotojen tuottamisessa. Twol-ohjelman suomenkielinen sanakirja sisälsi tutkimuksen ajankohtana noin 37 000 perusmuotoa (Karlsson 1990).

Syöte: Sananmuoto, esimerkiksi lääkkeitä.

Tuloste: Syötetyn sanan perusmuoto, kuten lääke. Käyttötarpeen mukaan Twol voi tuottaa muunkinlaisia tulosteita, kuten perusmuodon lisäksi kieliopillisen analyysin tuloksen: lääke N PTV PL.

Twol-ohjelmassa on kaksi eri osaa: sanakirja ja sääntökokoelma. Sanakirja sisältää sananvartalot sekä erilaiset etu- ja jälkiliitteet. Kaksitasomallin periaatteiden mukaisesti Twol-ohjelman toiminta perustuu kahteen eri tasoon: leksikaaliseen tasoon, joka on ohjelman sisäinen sanakirjamuoto, sekä pintatasoon, joka on sananmuodon todellinen esiintymä. Sananmuotojen analyysissä ja tuottamisessa Twol tutkii kaksitasosääntöjen avulla, vastaavatko leksikaalisen tason ja pintatason esitystavat toisiaan.

Twol-ohjelman kaksitasomallin mukaisia tasoja voidaan havainnollistaa kirjoittamalla muodot kahteen riviin päällekkäin seuraavasti:

leksikaalinen taso:	l a s i I A
pintataso	l a s e j a

Erilaisten äännevaihteluiden yms. vuoksi sanakirjassa olevat sananvartalot ja liiteainekset eivät välttämättä esiinny sellaisenaan sanan pintamuodossa. Tässä esimerkissä leksikaalisen tason muoto lasiIA koostuu kolmesta osasta: sananvartalosta lasi, monikon tunnuksesta I sekä partitiivipäätteestä A. Leksikaalisen tason monikon I kuitenkin toteutuu pintatasolla j-äänteenä ja sen edessä oleva sananvartalon i muuttuu e:ksi. (Koskenniemi 1985a).

Twol-ohjelman analyysissä siis tutkitaan, mihin sanakirjan sanoihin voidaan soveltaa suomen kielen sääntöjä siten, että soveltamisen tuloksena saadaan analysoitava sananmuoto. Analyysi etenee vasemmalta oikealle ja kaikki teoreettiset mahdollisuudet tutkitaan rinnakkain. Analyysin tuloksena tulostetaan ne sanat, jotka toteuttavat annetut ehdot.

6 TUTKIMUSONGELMA JA TUTKIMUSHYPOTEESIT

Käsillä olevan tutkimuksen tutkimusongelma määriteltiin ensimmäisessä luvussa alustavasti seuraavasti:

> Jos yksi tai useampia suomen kielen morfologisia tulkintaohjelmia liitetään osaksi tiedonhakujärjestelmää, niin voidaanko näiden tulkintaohjelmien avulla parantaa hakujärjestelmän hakemiston ominaisuuksia sekä nostaa hakutulosten saanti- ja tarkkuusarvoja?

Tutkimuksen lähtökohtana on, että hakujärjestelmänä on käänteishakemistoon ja Boolen logiikkaan perustuva tiedonhakujärjestelmä. Lisäksi oletetaan, että hakujärjestelmään tallennetut dokumentit ovat vapaamuotoista, aihealueeltaan rajoittamatonta tekstiä.

Perinteisten hakujärjestelmien käyttäjän on tiedettävä, miten sanat taipuvat, ja otettava tämä huomioon kyselyä tehdessään. Ammattilaishakija tämän yleensä osaakin, mutta satunnainen hakija ei välttämättä osaa käyttää hakujärjestelmän tarjoamia välineitä hyväkseen. Hän myös tekee ammattilaishakijaa useammin virheitä, joiden takia haluttuja dokumentteja jää löytymättä, esimerkiksi katkaisee hakusanat väärästä kohdasta tai jopa unohtaa katkaisun kokonaan.

Tiedonhakua helpottaisi, jos hakujärjestelmien käyttäjän, olipa hän sitten ammattilainen tai satunnainen hakija, ei tarvitsisi pohtia sanojen taipumista yms. esiintymätason ongelmia. Tämä voidaan toteuttaa periaatteessa kahdella eri tavalla:

- hakusanat katkaistaan automaattisesti
- hakemistoon tallennettavat sananmuodot normalisoidaan eli palautetaan perusmuotoon, jolloin hakusanoja ei tarvitse katkaista; hakuvaiheessa täsmäytetään perusmuotoiset hakusanat perusmuotohakemiston hakemistosanoihin

Hakusanojen automaattinen katkaisu toteutaan vartalo-ohjelmilla, jotka tuottavat perusmuodossa syötetystä sanasta sen vartalot (ks. luku 5.2.1). Vartalo-ohjelman tuottamien vartaloiden perään liitetään hakujärjestelmän jokerimerkki ja ne sijoitetaan kyselyyn alkuperäisen hakusanan sijasta -sulkujen sisällä, yhdistettynä toisiinsa TAI-operaattorilla. Tässä tapauksessa kysely tehdään taivutusmuotohakemistosta.

Jälkimmäisessä vaihtoehdossa puolestaan normalisoidaan taivutusmuodossa olevat merkkijonot (sananmuodot) perusmuodoiksi perusmuoto-ohjelmien avulla (ks. luku 5.2.2).

Tutkimusongelma tarkennetaan seuraaviksi seitsemäksi hypoteesiksi.

6.1 HAKEMISTOJEN KOKO

> **Hypoteesi 1**: Kun tekstin sananmuodot palautetaan perusmuotoon ennen kuin ne tallennetaan hakemistoon, perusmuotohakemistoon tulee vähemmän erilaisia merkkijonoja kuin vastaavasta tekstistä tuotettuun taivutusmuotohakemistoon. Koska hakemiston merkkijonojen määrä vähenee, perusmuotohakemisto vie kilotavuissa mitaten vähemmän muistitilaa kuin samasta tekstistä tuotettu taivutusmuotohakemisto.

Perusmuotoon palauttamisen voidaan olettaa vaikuttavan hakemistoon kahdella, toisilleen vastakkaisella tavalla: Toisaalta hakemiston merkkijonojen määrä vähenee, kun tietyn sanan kaikki sananmuodot yhdistetään yhteen perusmuotoon ja vain tämä tallennetaan hakemistoon edustamaan kaikkia esiintymiään. Toisaalta homografisista eli monitulkintaisista sananmuodoista löydetään enemmän kuin yksi tulkinta, jotka kaikki on otettava mukaan hakemistoon. Tämä lisää erityisesti hakemiston osoitteiden määrää ja näiltä osin muistitilan tarvetta. Lisäksi on otettava huomioon myös se, että perusmuoto-ohjelmat eivät tunnista kaikkia sananmuotoja. Tunnistamatta jääneet ilmaukset joudutaan tallentamaan taivutusmuotoisina. Näiden sananmuotojen osalta perusmuotohakemiston muistitilan tarve ei ole ainakaan pienempi kuin taivutusmuotohakemistossa.

Tosin on huomattava, että hakemiston tilankäyttöön vaikuttavat monet muutkin seikat kuin hakemistosanojen ominaisuudet. Tällaisia seikkoja ovat esimerkiksi osoitetarkkuus, hakemiston toteutustapa ja muistinvarausmenetelmä[1]. Vaikka perusmuoto-ohjelmien avulla voidaankin tuottaa erilaisia hakemistoversioita, jolloin erityyppiset hakemistot sisältävät erilaisia merkkijonoja, muut muistitilan käyttöön vaikuttavat tekijät saattavat kuitenkin vaimentaa hakemistojen välisiä eroja.

---

[1] Osoitetarkkuus - tallennetaanko osoitteet dokumentin, kappaleen, virkkeen vai sanan tarkkuudella; hakemiston toteutustapa - esimerkiksi β-puurakenteena; muistinvaraus - muistiavaruutta ei käsitellä yhtenäisenä kokonaisuutena, vaan se jaotellaan halutunkokoisiin lohkoihin tietojenkäsittelyn tehostamiseksi jne. (Ullman 1988)

Hypoteesille 1 löytyy tukea muun muassa Juha Niemistön (1988, s. 39 - 40) tutkimuksesta. Niemistö raportoi sanojen ja osoitteiden lukumäärien vaihtelut erilaisissa hakemistoissa, muttei näiden vaihteluiden yhteisvaikutusta hakemistojen muistitilan tarpeeseen esimerkiksi kilotavuina mitaten. Koska perusmuotoistaminen vaikuttaa hakemiston merkkijonojen ja osoitteiden määrään vastakkaisesti, on mielenkiintoista selvittää myös perusmuotoistamisen kokonaisvaikutus.

Perusmuotoistamisen vaikutusta tutkittiin FULLTEXT-projektissa varsinaisesti vertaamalla taivutusmuotohakemiston merkkijonojen ja osoitteiden määrää perusmuotohakemiston ja ositetun perusmuotohakemiston vastaaviin määriin (mukaanlukien kahteen viimeksimainittuun niitä täydentävä tunnistamattomien sananmuotojen hakemisto). Lisäksi tutkittiin näiden vertailtujen hakemistojen tilankäyttöä kilotavuissa mitaten, jotta saataisiin jonkinlainen näkemys merkkijonojen ja osoitteiden määrien muuttumisen kokonaisvaikutuksesta.

Vaikka tällä tutkimusasettelulla ei saadakaan absoluuttista kuvaa minkä tahansa tyyppisen hakemiston tilantarpeen muuttumisesta tarkkoina kilotavumäärinä (koska tämä riippuu kulloisenkin hakemistotyypin teknisestä rakenteesta) se kuitenkin havainnollistaa perusmuotoistamisen vaikutusta suhteessa muihin vaihtoehtoisiin tapoihin käsitellä sananmuotoja. Tässä tutkimusasetelmassahan kaikki hakemistot oli toteutettu samalla BASIS-hakujärjestelmällä ja muuten teknisesti samalla tavalla, paitsi että morfologisia tulkintaohjelmia sovellettiin vaihtoehtoisin tavoin. Näin ollen hakemistoon tulevien merkkijonojen ja osoitteiden määrä riippuu tekstin ja perusmuotoohjelmien ominaisuuksista ja on sinänsä riippumaton hakemiston teknisestä rakenteesta.

Hakemistojen tilantarve oli perusteltua tutkia FULLTEXT-projektissa myös siksi, että käytettävissä oleva aineisto oli noin 15-kertainen verrattuna Niemistön aineistoon: Niemistön käyttämä sanomalehtiaineisto sisälsi yhteensä 237 887 sananmuotoa, kun taas tämän tutkimuksen sanomalehtiaineistossa sananmuotoja oli yhteensä 3 582 356 kappaletta.

6.2 TAIVUTUSMUOTOHAKEMISTON HAKUOMINAISUUDET

Hypoteesit 2 ja 3 koskevat taivutusmuotohakemistosta tehtävää kyselyä:

> **Hypoteesi 2**: Kun ilmaisutason hakusanat esiintymätasolla korvataan vartalo-ohjelman hakusanoista tuottamilla taivutusvartaloilla, näillä automaattisesti katkaistuilla hakusanoilla saadun tulosjoukon tarkkuus on keskimäärin parempi ja saanti keskimäärin huonompi kuin tulosjoukon, joka on saatu hakijan katkaisemilla hakusanoilla.
>
> **Hypoteesi 3**: Kun ilmaisutason hakusanat esiintymätasolla korvataan vartalo-ohjelman hakusanoista sekä näiden johdosperheistä tuottamilla taivutusvartaloilla, näillä automaattisesti katkaistuilla hakusanoilla saadun tulosjoukon tarkkuus on keskimäärin parempi ja saanti keskimäärin sama kuin tulosjoukon, joka on saatu hakijan katkaisemilla hakusanoilla.

Hakusanojen automaattista katkaisua perustellaan usein sillä, että erityisesti tottumattomat käyttäjät saattavat katkaista hakusanat väärästä kohdasta, jolloin haluttuja dokumentteja voi jäädä löytymättä. Koska vartaloita tuottava ohjelma katkaisee hakusanat varmasti kieliopillisesti oikein, dokumentteja ei jää löytymättä tai löydy liikaa ainakaan katkaisuvirheiden takia. Automaattisesti katkaistuja hakusanoja käytettäessä tulosjoukon saannin pitäisi siis olla vähintään yhtä hyvä ja mahdollisesti parempikin kuin silloin, kun kyselyn hakusanat on katkaissut erehtyväinen ihminen. Hypoteesi 2 näyttää kuitenkin väittävän täsmälleen päinvastaista, kun siinä väitetään saannin huononevan verrattuna vastaavanlaiseen hakijan tekemään kyselyyn. Hypoteesin tarkoituksena ei kuitenkaan ole kiistää edellä esitettyä ajatuskulkua, vaan paremminkin astua siitä vielä askel eteenpäin.

Hypoteesin 2 väite tarkkuuden parantumisesta perustuu sille oletukselle, että vartalo-ohjelman tuottamat taivutusvartalot ovat yleensä pitempiä kuin asiantuntevan tiedonhakijan katkaisemat hakusanat. Harva hakija nimittäin on perehtynyt suomen kielioppiin niin hyvin, että osaisi vartalo-ohjelmien tapaan liittää vartaloiden loppuun vielä osia taivutuspäätteistä siten, että kaikki taivutusmuodot kuitenkin tulevat varmasti mukaan. Jos hakusana on katkaistu liian lyhyeksi, se helposti palauttaa hakemistosta myös vääriä sananmuotoja, jotka vain sattumalta alkavat samalla merkkijonolla kuin hakusana. Mitä lähempää sanan loppua hakusana katkaistaan, sitä enemmän tällaisia vääriä osumia jää pois ja siten tulosjoukon tarkkuus kasvaa.

Hypoteesin 2 väite saannin huononemisesta perustuu samaan asiaan: pitkät taivutusvartalot voivat karsia pois myös hyödyllisiä hakemiston merkkijonoja. Merkitykseltään läheiset sanat ovat usein myös merkkijonoina lähellä toisiaan; tästä tyypillinen esimerkki ovat johdokset. Jos hakusanan ja sen johdoksen yhteinen alkuosa on kovin lyhyt eli käytännössä lyhyempi kuin hakusanan taivutusvartalo, taivutusvartalo täsmää kyllä hakemistosta löytyvien hakusanan esiintymien kanssa, mutta johdosten esiintymät karsiutuvat pois. Tämän vuoksi taivutusvartaloilla saadun hakutuloksen saanti jää huonommaksi kuin hakijan katkaisemia hakusanoja käytettäessä.

Hypoteesin 3 mukaan edellä kuvattu ongelma voidaan välttää, kun hakusanojen johdoksetkin otetaan kyselyssä huomioon. Vartalo-ohjelmalle syötetään sekä itse hakusana että sen johdosperheen jäsenet ja näin saadut vartalot lisätään kyselyyn. Kun näin saadaan johdoksetkin haetuiksi, tulosjoukon saantiarvo pysyy samana kuin hakijan itse katkaisemia hakusanoja käytettäessä. Toisaalta tulosjoukon tarkkuusarvon pitäisi edelleenkin olla hyvä eli parempi kuin perinteisellä tavalla haettaessa. Ovathan sekä itse hakusanan että sen johdosperheen jäsenten taivutusvartalot yleensä pidempiä kuin hakijan itse katkaisemat hakusanat ja siten tuottavat vähemmän epärelevantteja dokumentteja tulosjoukkoon.

6.3 PERUSMUOTOHAKEMISTON HAKUOMINAISUUDET

> **Hypoteesi 4**: Kun dokumenttien sisältämät sananmuodot perusmuotoistetaan sekä tallennetaan perusmuotoisina hakemistoon ja myös hakija syöttää hakusanat perusmuodossa, näillä perusmuodoilla perusmuotohakemistosta saadun tulosjoukon tarkkuus on keskimäärin parempi mutta saanti keskimäärin huonompi kuin tulosjoukon, joka on saatu taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla.

Tutkimushypoteesit 2 ja 4 kuvaavat samaa ilmiötä erityypisissä hakemistoissa: Hypoteesin 2 tapauksessa kysely kohdistuu taivutusmuotohakemistoon, hypoteesin 4 tapauksessa perusmuotohakemistoon. Kummassakin tapauksessa hakija syöttää hakusanat perusmuotoisina, mutta hypoteesin 2 tapauksessa (ilmaisutason) hakusanat syötetään ensin vartalo-ohjelmalle, jonka tuottamat vartalot sijoitetaan hakusanan tilalle (esiintymätason) kyse-

lyssä, hypoteesissa 4 taas perusmuotoja käytetään sellaisinaan hakusanoina perusmuotohakemistosta haettaessa[2].

Hypoteesissa 2 hakusanat katkaistaan automaattisesti, jolloin tuloksena on pidempiä ja siten täsmällisempiä merkkijonokaavioita kuin silloin, kun hakija katkaisee hakusanat itse. Hypoteesin 4 tapauksessa taas perusmuotoiset hakusanat täsmäytetään suoraan hakemistosanoihin eli voidaan hakea suoraan merkkijonovakioilla - näitä vakioita yleisempiä ja samalla epätarkempia merkkijonokaavioita ei tarvitse käyttää.

Kun perusmuotoisia hakusanoja ja hakemistosanoja verrataan suoraan toisiinsa, hakusanojen kanssa täsmäävät vain niiden kanssa merkki merkiltä täysin identtiset hakemistosanat. Perusmuotohakemistosta saadun tulosjoukon tarkkuus on hyvä, koska näin siihen ei joudu vain sattumalta hakusanan kanssa samalla merkkijonolla alkavia sananmuotoja.[3] Toisaalta saanti kärsii, kun hakusanoille merkitykseltään läheiset, mutta merkkijonoina erilaiset sanat jäävät pois; tällaisia ovat esimerkiksi johdokset. Jos hakija siis haluaa alkuperäisten hakusanojen lisäksi ottaa huomioon myös niille merkitykseltään läheiset johdokset ja yhdyssanat, on hänen varta vasten lisättävä nämä kyselyyn (johdokset -> hypoteesi 5, yhdyssanat -> hypoteesi 6).

> **Hypoteesi 5**: Kun dokumenttien sananmuodot palautetaan perusmuotoon sekä tallennetaan perusmuotoisina hakemistoon, ja myös hakijan syöttämät hakusanat johdosperheineen ovat perusmuodossa, tällä johdosperheellä perusmuotohakemistosta saadun tulosjoukon tarkkuus on keskimäärin parempi, mutta saanti keskimäärin sama kuin tulosjoukon, joka on saatu taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla.

Hypoteesi 5 on vastaava laajennus hypoteesista 4 kuten hypoteesi 3 oli hypoteesin 2 laajennus. Kun kyselyssä käytetään täsmällisiä perusmuotoja, hakusanoille merkityssisällöltään läheiset johdokset jäävät tulosjoukosta

---

[2] Tosin käytännössä myös jälkimmäisessä vaihtoehdossa tarvitaan vartalo-ohjelmia, koska haku on perusmuotohakemiston lisäksi kohdistettava myös tunnistamattomien sananmuotojen hakemistoon. Koska sananmuodot on tallennettu tunnistamattomien sananmuotojen hakemistoon taivutusmuotoisina, niitä on haettava katkaistuilla hakusanoilla.

[3] Tosin väärintulkitut (sananmuoto)homografit voivat tuottaa tulosjoukkoon epärelevantteja dokumentteja, koska kaikki perusmuoto-ohjelman tuottamat tulkinnat on tallennettava hakemistoon. Esimerkiksi nimi Salmén voidaan tulkita salmi-sanan genetiivimuodoksi. Tästä homografiongelmasta enemmän luvussa 10.

pois. Kun johdokset lisätään kyselyyn mukaan, hakutuloksen saannin pitäisi nousta samalle tasolle kuin hakijan katkaisemilla hakusanoilla haettaessa. Koska perusmuotoiset hakusanat ovat täsmällisiä eivätkä tuota tulosjoukkoon ylimääräisiä, epärelevantteja dokumentteja, niillä saadun tulosjoukon tarkkuus on parempi kuin tulosjoukolla, joka on saatu hakijan katkaisemilla hakusanoilla.

> **Hypoteesi 6**: Kun hakija tekee perusmuotohakemistosta kyselyn, jossa hakusanat ja hakusanojen johdosperheen jäsenet ovat perusmuodossa ja lisäksi hakusana ja tämän johdosperheen jäsenet katkaistaan automaattisesti näillä sanoilla alkavien yhdyssanojen löytämiseksi, tällaisen kyselyn tuloksena saadun tulosjoukon tarkkuus on keskimäärin parempi ja saanti keskimäärin parempi kuin tulosjoukon, joka on saatu taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla.

Hypoteesi 6 esittää, että kun edellisen hypoteesin 5 esittämää tilannetta jatketaan laajentamalla kyselyä hakusanan ja sen johdosperheen sisältäviin yhdyssanoihin, tulosjoukon saanti paranee, mutta kuitenkin niin, ettei tarkkuus huonone samassa suhteessa (yhtä monta prosenttiyksikköä) kuin saanti samalla paranee. Koska hakemistosanat perusmuotohakemistossa ovat normaalistettuja sanoja eivätkä taivutusmuotoja, tulosjoukkoon ei eksy niin paljon epärelevantteja dokumentteja kuin taivutusmuotohakemistossa.

6.4 OSITETUN PERUSMUOTOHAKEMISTON HAKUOMINAISUUDET

> **Hypoteesi 7**: Kun hakija tekee ositetusta perusmuotohakemistosta kyselyn, jossa hakusanat, hakusanojen johdosperheen jäsenet ja yhdyssanojen osat ovat perusmuodossa ja yhdyssanan osina, tällaisen kyselyn tuloksena saadun tulosjoukon tarkkuus on keskimäärin huonompi, mutta saanti keskimäärin parempi kuin tulosjoukon, joka on saatu taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla; lisäksi tällaisen kyselyn saanti on keskimäärin parempi kuin (osittamattomasta) perusmuotohakemistosta saadun tulosjoukon saanti.

Hypoteesissa 7 lähtökohtana on, että perusmuotoon palauttamisen yhteydessä on myös jaettu yhdyssanat osiinsa. Näin hakemistoon on tallennettu erillisinä sanoina myös yhdyssanan alussa, keskellä ja lopussa esiintyneet sanat, mahdollisesti myös niiden yhdistelmät. Perinteisestä taivutusmuotohakemistosta löydetään helposti vain yhdyssanan alkuosat; yhdyssanan keski- ja loppuosat jäävät löytymättä, ellei hakija keksi kaikkia alkuosia, jotka ovat esiintyneet yhdyssanassa hakusanojen edellä. Monissa haku-

järjestelmissä tosin on mahdollista katkaista hakusanat myös vasemmalta, jolloin sanojen keski- ja loppuosatkin ovat löydettävissä. (Yhdyssanojen osittamisen ongelmista tarkemmin luvussa 7.2.)

Kun hakusanat poimivat ositetusta perusmuotohakemistosta myös sellaiset sanat, joiden keskellä tai lopussa hakusana on esiintynyt, kysely tuottaa tulosjoukkoon dokumentteja, joihin taivutusmuotohakemistossa tai pelkässä perusmuotohakemistossa ei päästä käsiksi. Näin saadun tulosjoukon saannin siis pitäisi olla parempi kuin vastaavan, hakijan katkaisemilla hakusanoilla saadun tulosjoukon. Toisaalta hakusanan kanssa voivat täsmätä sellaisetkin yhdyssanat, joiden merkitys voi olla melko etäällä hakijan alunperin antaman hakusanan merkityksestä. Niinpä yhdyssanojen lisäämisestä kyselyyn voi seurata, että tulosjoukon tarkkuusarvo huononee.

Hypoteeseille 2 - 7 ei voida esittää aiemman kirjallisuuden perusteella tukea samalla tavoin kuin hypoteesille 1, koska tällaista kirjallisuutta ei tutkimuksen suunnitteluvaiheessa hypoteeseja laadittaessa ollut käytettävissä. Hypoteesien pohjana on järkeily, jossa perusteena ovat olleet morfologisten tulkintaohjelmien toimintaperiaatteet sekä aiempien julkaisujen teoreettiset hahmotelmat ja suppeammat kokeilut (Nurminen 1986, Hjorth 1987b, Niemistö 1988). Luvussa 4 esitelty englannin kielen karsinta-algoritmien tutkimus ei ole sellaisenaan hyödynnettävissä suomenkielistä aineistoa tutkittaessa - etenkään yhdyssanojen osalta.

6.5 VAKIO- JA ONGELMAKYSELYT

Täsmennyksenä hypoteeseille todettakoon vielä, että ne liittyvät **vakiokyselyihin**, joissa hakusanat ovat perusmuoto-ohjelmien sanakirjassa esiintyviä sanoja eivätkä edellytä hakijalta tai hakujärjestelmältä lisätarkistuksia tai erityiskäsittelyjä. Näissä tapauksissa hakusanat ja hakemistosanat käyttäytyvät morfologisten ohjelmien olettamalla tavalla ja siten näiden ohjelmien kielioppisäännöt toimivat kuten on tarkoituskin.

Käytännössä kuitenkin törmätään myös sellaisiin ilmauksiin, joissa esiintyy esimerkiksi perusmuoto-ohjelmille tuntemattomia sanoja tai taivutusmuotoisia sananmuotoja. Näissä **ongelmakyselyissä** eivät edellä esitetyt hypoteesit toimi kuten olisi tarkoitus.

Vakio- ja ongelmakyselyjen suhteellisesta osuudesta ei ole täsmällistä tietoa, mutta todennäköistä on, että ongelmakyselyjen lukumäärä korreloi tallennusvaiheessa tunnistamatta jäävien sananmuotojen määrän kanssa. Tällä perusteella voidaan arvioida, että valtaosa kyselyistä olisi vakiokyselyjä.

Vaikka tämä tutkimus keskittyy vakiokyselyihin ja niiden käyttäytymiseen erityyppisissä hakemistoissa, hypoteesien testausta täydennettiin erillisellä osuudella, jossa analysoitiin erilaisia ongelmakyselytyyppejä sekä suunniteltiin ratkaisuja, joilla tällaisetkin kyselyt voitaisiin suorittaa (ositetusta) perusmuotohakemistosta. Ongelmakyselyjen erityispiirteitä tarkastellaan luvussa 10.

7 TUTKIMUKSEN TOTEUTUS

Tutkimushypoteesejä tutkittiin FULLTEXT-projektissa kokeellisesti rakentamalla erilaisia testijärjestelmiä. Pyrkimyksenä oli muuttaa yhtä vaikuttavaa tekijää kerrallaan ja tulosten perusteella arvioida eri tulkintaohjelmien ja niiden eri soveltamistapojen vaikutus tiedonhaun tuloksiin. Vaikka morfologisten tulkintaohjelmien soveltamisperiaatteita onkin teoreettisesti hahmoteltu muutamissa aiemmissa tutkimuksissa (Nurminen 1986, Hjorth 1987b, Niemistö 1988), vastaavanlaajuista vertailua ei käytännössä ollut aiemmin tehty. Erilaiset testijärjestelmät oli FULLTEXT-projektissa suunniteltava ja toteutettava itse ilman varsinaisia aiempia esikuvia. Tässä luvussa kuvataan tarkemmin tutkimusaineisto ja tutkimusympäristöjen toteutustapa.

Tutkimuksen toteutus jakautui seuraaviin vaiheisiin:

1. tekstiaineiston hankinta (katso luku 7.1)
2. eri hakemistovaihtoehtojen arviointi erityisesti yhdyssanojen osittamisen kannalta (luku 7.2)
3. hakemistojen tuottaminen (luku 7.3)
4. varsinaisten tutkimusympäristöjen rakentaminen (luku 7.4)
5. virheenkorjausmenetelmien suunnittelu ja toteutus (luku 7.5)
6. hakupyyntöjen keräys ja alustava valikointi sekä karsinta koekyselyjen perusteella (luku 7.6.1)
7. hakusanojen poimiminen hakupyynnöistä (luku 7.6.2)
8. kyselytyyppien muodostaminen eri tutkimusympäristöjen vertailua varten (luku 7.6.3)
9. hakutulosten relevanssin arviointi (luku 7.7)
10. saannin ja tarkkuuden laskeminen (luku 7.8)
11. otantaperiaatteet (luku 7.9)
12. merkitsevyystestit (luku 7.10).

7.1 Tekstiaineiston hankinta

Tutkimuksen testiaineistona oli sanomalehden toimituksellinen teksti. Sanomalehden sisällöstä toimituksellista tekstiä on pinta-alan perusteella mitaten

noin puolet. Muu aineisto on ilmoituksia ja toimituksellisia kuvia (Lehti luukussa... 1989, s. 56). Tyypillisesti sanomalehtiteksti koostuu monentyyppisistä artikkeleista: niin uutisista, reportaaseista, kulttuuriarvosteluista ja yleisönosastokirjoituksista kuin urheilun tulosluetteloista, tapahtumakalentereista ja pörssikursseistakin. Jutut ovat rakenteeltaan ja pituudeltaan hyvin vaihtelevia. Tyypillisesti sanomalehti elää oman julkaisupäivänsä tilanteessa. Artikkelin kirjoittaja ei välttämättä ota huomioon, että artikkeli voidaan haluta lukea vielä pitkänkin ajan kuluttua; ensisijainen tavoite on saada artikkeli kirjoitetuksi määräaikaan mennessä. (Hjorth 1987b)

Sanomalehtitekstissä voi olla kirjoitusvirheitä, koska niiden kaikkien etsimiseen ja korjaamiseen ei yleensä ole aikaa tai se tulisi liian kalliiksi. Lehtiteksti sisältää paljon nimiä, nimilyhenteitä ja eri tavoin kirjoitettuja numerotietoja. Erityisesti ulkomaiset nimet ovat suuri ja jatkuvasti kasvava joukko. Monesti sanomalehdet kertovat ensimmäisenä uusista asioista, jolloin uusien käsitteiden ja nimien kirjoitustapa voi vaihdella huomattavasti. Myöhemmin, asian yleistyttyä tai jäätyä historiaan, ei tiedonhakija välttämättä osaa palauttaa mieleensä kaikkia vaihtoehtoja (esimerkiksi EEC, EC, EY, EL, EU). Joistain asioista taas voi olla jatkuvasti käytössä useita nimityksiä tai kirjoitustapoja rinnakkain. (Hjorth 1987b) Tiukkatahtisessa, aikatauluun ja rajallisiin resursseihin sidotussa lehtityöskentelyssä elektronisen sanomalehtiarkiston erityistarpeita ei siis välttämättä oteta huomioon.

FULLTEXT-projektin testiaineistona oli kolmen kuukauden otos Aamulehden elektronisesta tekstiarkistosta. Kaikkiaan dokumentteja kertyi tutkimustietokantaan 23 244 kappaletta. Kokonaisuudessaan Aamulehden arkistojärjestelmä sisältää lehden artikkelit heinäkuusta 1989 alkaen, mikä FULLTEXT-projektia aloitettaessa oli kaikkiaan satojatuhansia artikkeleita. Aamulehden arkistojärjestelmään tallennettiin Aamulehdessä julkaistusta 200 - 300 päivittäisestä artikkelista noin 180 - 240 artikkelia. Tallentamatta jätettiin mm. pörssikurssit ja radio-ohjelmat. Tekstiarkiston vuosittainen kertymä oli tutkimusaineiston keräämisen aikana noin 70 000 artikkelia. Artikkelin tyypillinen pituus oli noin 3 000 - 5 000 merkkiä. (Ylinen 1991)

7.2 Yhdyssanan osiin jakamisen ja tallentamisen vaihtoehdot

FULLTEXT-projektissa haluttiin tutkia erityisesti yhdyssanojen erilaisia tallennus- ja hakuvaihtoehtoja. Koska perusmuoto-ohjelmat voivat haluttaessa purkaa yhdyssanat osiinsa, voidaan kehittää uudemmanlaisia hakemistoratkaisuja kuin perinteisesti on ollut mahdollista. Esimerkiksi yhdys-

sanojen keski- ja loppuosia on hankala hakea taivutusmuotohakemistosta. Jos hakija siinä haluaa yhdyssanojen saannin olevan mahdollisimman hyvän, hänen on osattava keksiä kaikki mahdolliset määriteosat, jotka halutun sanan edellä saattavat esiintyä, kuten sokeri -> hedelmäsokeri, ruokosokeri, koivusokeri jne. Yhdyssanaongelmaa on perinteisissä tiedonhakujärjestelmissä yritetty ratkaista teknisesti, muun muassa sallimalla hakusanan katkaisu vasemmalta (**sokeri*) tai tallentamalla sananmuodot hakemistoon myös takaperin (irekosämledeh), taikka käsitteellisesti tesaurusten avulla, jolloin yläkäsitteen (kuten sokeri-perussanan) kautta haetaan sen alakäsitteitä (sokeri-loppuiset yhdyssanat) ja päinvastoin.

Kaikki hakujärjestelmät eivät salli vasemmalta katkaisua, koska se kuluttaa paljon tietokoneresursseja. Tämä ongelma voidaan kuitenkin ratkaista hakemistolla, joka sisältää sananmuodot takaperin, koska tässä tapauksessa sanan loppuosan haku ei ole sen raskaampaa kuin tavallinen haku (*irekos**). Käytännössä kumpikin näistä teknisistä ratkaisutavoista tuottaa ongelmia suomenkielisen aineiston käsittelemisessä: perusmuotoista hakusanaa käytettäessä jäävät yhdyssanan taivutusmuodot (esimerkiksi hedelmäsokerilla tai allirekosämledeh) löytymättä. Hakusana on siis katkaistava molemmin puolin, eli **sokeri** tai **irekos**.

Molemminpuolisen katkaisun haittana kuitenkin on, että jos hakusana on lyhyt, se voi tuottaa paljon epärelevantteja dokumentteja, joissa esiintyy pidempiin sanoihin sisältyviä **loissanoja**. Loissanat muodostuvat merkkijonoista, jotka sinänsä ovat todellisia sananmuotoja, mutta eivät kyseisessä tapauksessa ole sanan varsinaisia osia, kuten sanomassa -> sanoma, sano, anomassa, ano, oma, omassa, massa.

Tesaurusten ongelmana taas on, että niiden laatiminen vaatii paljon resursseja. Jos jokaisen perussanan alle listattaisiin kaikki siitä muodostuvat yhdyssanat, tesauruksesta tulisi erittäin laaja. Käytännössä tuskin pystytään siihen, että tesaurus sisältäisi kaikki mahdolliset yhdyssanat, ainakaan jos tesaurus laaditaan manuaalisesti.

Kun perusmuotoon palauttavat ohjelmat jakavat yhdyssanat osiinsa, hakemistoon voidaan haluttaessa tallentaa yhdyssanan lisäksi myös sen osat sekä erilaiset osien yhdistelmät. Tämä on käytännössä hakusanan vasemmalta ja oikealta katkaisun erikoistapaus. Tässä tapauksessa katkaisu vain tehdään jo ennakolta tallennusvaiheessa ja kieliopillisin perustein, jolloin loissanoilta periaatteessa vältytään. Käytännössä loissanoja kuitenkin tulee hakemistoon, koska tekstiyhteydestään irralliset sananmuodot voivat olla monitul-

kintaisia (esimerkiksi veronalainen -> veron-alainen vai verona-lainen; syyssään -> syys-sään vai syy-ssä-än). Kaikki tulkinnat on tallennettava hakemistoon, vaikka vain yksi niistä onkin oikea.

Kun FULLTEXT-projektia varten suunniteltiin eri hakemistovaihtoehtoja, suunnittelussa käytettiin muun muassa seuraavia arviointiperusteita. Sopivien esikuvien puuttuessa tutkija laati ne projektia varten itse:

1. Kyselyn laajentaminen ja supistaminen. Kyselyn alaa voidaan laajentaa lisäämällä yhdyssanan osat mukaan kyselyyn vaihtoehtoisiksi hakusanoiksi. Toisaalta hakijan pitää voida rajata kysely mahdollisimman täsmälliseen ilmaukseen. Perussana ei yleensä ole yhtä täsmällinen kuin siitä määriteosalla muodostettu yhdyssana. Tosin yhdyssanan osa voi olla itsenäisenä esiintyessään spesifi ilmaus, jonka merkitys on eri kuin yhdyssanan osana. Yhdyssanan osat pitää voida erottaa itsenäisinä esiintyneistä perussanoista (kunta - seurakunta, kasvikunta, eduskunta).

2. Yhdyssanajaon ymmärrettävyys. Hakijan on ymmärrettävä yhdyssanojen osittamisen yleisperiaatteet, jotta hän pystyy antamaan hakusanat sopivassa muodossa sekä laajentamaan tai supistamaan haun alaa. Mitä mutkikkaammin sanat on tallennusvaiheessa käsitelty, sitä enemmän hakuvaiheessa tarvitaan opastusta. Hakijalta ei pidä edellyttää, että hänen on täysin hallittava yhdyssana-analyysin tekninen toteutus ja se, millä tavalla kukin sananmuoto tulee tulkituksi, vaan nämä asiat on piilotettava hakujärjestelmän käyttöliittymän taakse.

   Hakujärjestelmän on käsiteltävä hakijan antama syöte samojen periaatteiden mukaisesti (= tulkintaohjelman samalla versiolla ja sanakirjalla) kuin sananmuodot käsiteltiin tallennusvaiheessa, jotta kyselyn hakusanat olisivat yhteismitallisia hakemistosanojen kanssa. Mutta loppujen lopuksi kuitenkin hakijan itsensä on osattava päätellä, miten yhdyssanojen osat ovat vaikuttaneet hakutuloksiin, jotta hän halutessaan kykenee muokkaamaan kyselyä.

3. Hakemiston selattavuus. Hakijan on helpompi ideoida uusia hakusanoja, jos hänellä on mahdollisuus selata käänteistiedostoa ja poimia sieltä dokumenteissa esiintyneitä ilmauksia. Yhdyssanat (kuten lentokone ja ydinvoima) ovat selkeämmin tunnistettavia ilmauksia kuin niiden eri puolilla hakemistoa olevat osat (lento- ja -kone, ydin- ja -voima).

4. Hakemiston osoitetarkkuus. Vaikka yhdyssanan osat olisivat erillään hakemistossa, ne on hakuvaiheessa yksinkertaista yhdistää toisiinsa, mikäli osoitteet ovat hakemistossa sanan tarkkuudella. Todennäköisyys, että hakemistosta löydetyt osat kuuluvatkin eri sanoihin on sitä suurempi, mitä epätarkempia (virke, kappale, dokumentti tms.) osoitteet ovat. Mikäli osoitetarkkuus on huono, yhdyssanat eivät voi olla

hakemistossa pelkästään osinaan, vaan myös kokonaisen yhdyssanan on löydyttävä hakemistosta.

5. Hakemiston tarvitsema muistitila. (Yhdys)sanojen eri taivutusmuotojen perusmuotoistaminen vähentää hakemistoon tulevien sanojen määrää, mutta yhdyssanojen osien lisääminen taas lisää hakemiston muistitilan tarvetta. Eri jakamistavat tuottavat eri määrän yhdyssanojen osia eli kuormittavat hakemistoa eri tavoin.

6. Analyysivirheiden korjaaminen. Perusmuotoon palauttamisessa voi sattua virheitä - esimerkiksi silloin, kun sananmuotohomografin perusmuoto puuttuu sanakirjasta. Mikäli sanakirjasta puuttuu Alkula, sanamuoto Alkulakin tulkitaan yhdyssanaksi alku + lakki (edellyttäen, että nämä perusmuodot ovat sanakirjassa). Jos hakemistoon on tallennettu vain yhdyssanojen erilliset osat, on osien kokoaminen takaisin yhdeksi sanaksi tällaisten tulkintavirheiden tapauksessa hyvin työlästä; analyysivirheet on helpompi korjata, kun hakemistoon on tallennettu myös yhdyssanat kokonaisuudessaan (alkulakki).

7. Muiden, perinteisestä poikkeavien tallennus- ja hakumenetelmien tarpeet. Hakemistosanojen perusmuotoistaminen tekee mahdolliseksi menetelmät, jotka perinteisessä taivutusmuotohakemistossa ovat hankalia toteuttaa. Tällaisia "uusmenetelmiä", joita ulkomailla on tutkittu jo vuosia, mutta joita suomenkielisissä hakujärjestelmissä ei FULLTEXT-projektin ajankohtana ollut sovellettu, ovat esimerkiksi sanojen esiintymistaajuuteen perustuvien relevanssiarvojen laskeminen ja automaattinen indeksointi. Vaikka näitä menetelmiä ei itse projektissa tutkittukaan, siinä pohdittiin myös sitä, millainen yhdyssanojen käsittelytapa olisi sopivin, kun hakujärjestelmässä ei käytetä Boolen operaattoreita.

Edellä esitettyjen perusteiden avulla arvioitiin, millainen vaikutus eri yhdyssana-analyyseillä todennäköisesti olisi hakujärjestelmien toimintaan. Arvioiden perusteella päätettiin, mitkä mahdollisista hakemistovaihtoehdoista kannattaisi toteuttaa projektissa.

Suunnitelmavaiheessa myös laskettiin, mitä eri vaihtoehdoissa tapahtuu, kun perusmuoto-ohjelma käsittelee *n*-osaista yhdyssanaa. Oletuksena on, että yhdyssanan kaikki osat löytyvät perusmuoto-ohjelman sanakirjasta ja että löytyy vain yksi (oikea) tulkinta. Esimerkiksi sanassa ydinvoimalaitos $n = 3$ (ydin, voima, laitos).

Yksinkertaisin käsittelytapa on, että kaikki **hakusanat pelkästään palautetaan perusmuotoon** eikä yhdyssanoja jaeta osiinsa. Tällöin yhdyssanojen keski- ja loppuosia ei pystytä hakemaan (esimerkiksi ydinvoimalaitoksessa

-> ydinvoimalaitos), vaan haun laajentamiseen ja supistamiseen on samanlaiset mahdollisuudet kuin taivutusmuotohakemistossa. Koska yhdyssanat käsitellään aivan samalla tavalla kuin perussanat, analyysitavan ymmärrettävyys on hyvä. Osoitetarkkuudelle tämä tapa ei aseta erityisvaatimuksia. Perusmuotohakemistoa on selkeämpi selata kuin taivutusmuotohakemistoa, koska hakemisto rakentuu sanoista eikä sananmuodoista. Tämä vähentää myös hakemiston tarvitsemaa muistitilaa. Virhetulkintojen korjaamisen ja uusmenetelmien suhteen yhdyssanojen pelkkä perusmuotoistaminen ei aiheuttane erityisongelmia. FULLTEXT-projektissa tämä vaihtoehto toteutettiin, koska haluttiin nähdä, miten pelkkä taivutusmuotojen poisjättäminen hakemistosta ja näiden taivutusmuotojen osoitetietojen liittäminen perusmuotoon vaikuttaa hakemiston ominaisuuksiin ja hakutuloksiin (hakemistotyyppi H2; katso luku 7.3.3).

Toinen tapa rakentaa hakemisto on **hajottaa yhdyssanat osiinsa ja tallentaa vain nämä osat** hakemistoon. Hakuvaiheessa hakijan antama yhdyssana pilkotaan vastaavalla tavalla, osien väliin sijoitetaan mahdollisimman tiukka läheisyysoperaattori ja osat haetaan hakemistosta. Tämäkin käsittelytapa on hakijalle periaatteessa ymmärrettävä: yhdyssana on osiensa summa. Osien tallentaminen voidaan tehdä kahdella tavalla: joko tallennetaan sanat sinänsä tai sitten liitetään osiin myös sijaintitieto eli tieto siitä, missä kohtaa yhdyssanaa ne ovat esiintyneet.

Jos hakemistoon ei tallenneta sijaintitietoa, säästetään huomattavasti muistitilaa, mutta haun tarkkuus kärsii (ydinvoimalaitoksessa -> ydin, voima, laitos). Ensinnäkin tilaa säästyy, kun hakemistossa on eri taivutusmuotojen sijasta vain perusmuotoja. Oletettavasti yhdyssanan osat löytyvät hakemistosta muutenkin, itsenäisinä perussanoina, joten yhdyssanan jakaminen ei enää lisää hakemistosanojen määrää. Lisäksi suomenkielinen teksti sisältää niin paljon yhdyssanoja, että niiden karsimisen voi olettaa supistavan hakemiston kokoa huomattavasti. Esimerkiksi Niemistön (1988, s. 31) tutkimuksessa perusmuotoistamisen jälkeen noin 40 prosenttia sanoista oli yhdyssanoja. Tässä kuvatun lähestymistavan ongelmana kuitenkin on, että muistitilaa säästetään haun tarkkuuden kustannuksella. Muun muassa henkilönnimissä on syytä pitää itsenäiset sanat ja yhdyssanan osat erillään (Mäki, muttei Keskimäki, Mäenpää yms.). Huonon tarkkuuden vuoksi tätä vaihtoehtoa ei pidetty projektissa toteutuskelpoisena.

Mikäli yhdyssanojen jakaminen toteutetaan niin, että osiin merkitään niiden sijainti yhdyssanassa, haut pystytään tarkentamaan paremmin (ydinvoima-

laitoksessa -> ydin-, -voima-, -laitos). Yhdyssanojen kokoaminen osistaan kuitenkin edellyttää, että hakujärjestelmään on tallennettu riittävän tarkat osoitetiedot. BASIS-K-hakujärjestelmässä hakusanojen sijainti voitiin määritellä virkkeen tarkkuudella. Tämä ei aina riitä rajaamaan epärelevantteja dokumentteja pois. Virkeoperaattoria tai sitä väljempää operaattoria käytettäessä saataisiin esimerkiksi *liitukausi*-hakusanalla myös virke, jossa esiintyisivät samanaikaisesti sekä liitutaulu- että lukukausi-sanat. Tällaiset osumat karsiutuvat hakujärjestelmässä, jossa osoitteet ilmaistaan sanan tarkkuudella. Tällöin yhdyssanan osat saavat tallennusvaiheessa saman osoitteen ja hakuvaiheessa vain ne osat, joilla on yhteinen osoite, kytketään toisiinsa.

Kun yhdyssanan osiin merkitään sijaintitiedot, osoitteiden määrä hakemistossa kasvaa, kun osoiteviittaukset tarvitaan yhden sananmuodon sijasta yhdyssanan kaikille *n*:lle osalle. Lisäksi yhdyssanojen osina esiintyneistä sanoista generoituu uusia hakemistosanoja, periaatteessa kolme uutta esiintymää kunkin itsenäisen perussanan seuraksi (kuten ydin -> ydin-, -ydin-, -ydin; esimerkiksi ydinvoima, selkäydinneste, luuydin). Käytännössä kaikki sanat eivät kuitenkaan esiinny yhdyssanan kaikissa mahdollisissa asemissa, joten yhdyssanojen jakaminen vaikuttanee muistitilaan loppujen lopuksi suhteellisen vähän. Hakusanojen ideoinnin kannalta on puute, että hakemistoa selattaessa ei löydetä yhdyssanoja. Myös väärin tulkittujen sananmuotojen jäljittäminen hakuvaiheessa sekä sanojen esiintymistiheyden laskeminen on hankalaa. Näin ollen tätäkään vaihtoehtoa ei pidetty käyttökelpoisena projektin BASIS-toteutuksessa.

Seuraava vaihtoehto on tallentaa **hakemistoon sekä kokonaiset yhdyssanat että niiden osat**; osiin merkitään niiden sijainti yhdyssanassa (ydinvoimalaitoksessa -> ydinvoimalaitos, ydin-, -voima-, -laitos). Tämä tapa ratkaisee useimmat edellä esitetyt ongelmat. Yhdyssanojen keski- ja loppuosia voidaan hakea. Hakemistoa selattaessa nähdään sekä perussanat että yhdyssanat. Myös perusmuoto-ohjelman mahdollisesti tekemien väärien tulkintojen korjaaminen sekä sanojen esiintymistaajuuden laskeminen onnistuu edellistä vaihtoehtoa paremmin. Sanojen esiintymistaajuuden laskeminen on kuitenkin ongelmallista, jos yhdyssanassa on enemmän kuin kaksi osaa. Esimerkkitapauksessa ydinvoimalaitos ja sen kukin yksittäinen osa voidaan laskea, mutta yhdyssanaan myös sisältyvät ydinvoima- ja -voimalaitos jäävät pois laskelmista. Muistitilaa tarvitaan jonkin verran enemmän kuin edelli-

sessä vaihtoehdossa, koska yhdyssanasta generoituu 1+*n* sanaa ja 1+*n* osoitetta.

BASIS-K-hakujärjestelmän tapauksessa ongelmaksi jää vielä yhdyssanan osien kytkemisen tarkkuus: kun osoitteet ilmoitetaan virkkeen tarkkuudella, kytkettävät osat eivät välttämättä olekaan esiintyneet samassa sanassa. Tämän vuoksi FULLTEXT-projektin BASIS-osuudessa ei toteutettu tätä vaihtoehtoa. Sen sijaan projektin APL-Minttu-ympäristössä tällainen hakemisto rakennettiin, koska siinä osoitteet voitiin tallentaa hakemistoon yksittäisen sanan tarkkuudella (Alkula & Honkela 1992, s. 45).

Edellisen vaihtoehdon sijasta BASIS-hakujärjestelmässä toteutettiin vaihtoehto, jossa hakemistoon tallennettiin **yhdyssanojen ja niiden osien lisäksi vielä osien yhdistelmät** (hakemistotyyppi H3; luku 7.3.3). Kun yhdyssanojen osat kytketään toisiinsa, haun laajennettavuus on yhdyssanojen suhteen maksimaalinen. Perussanat ja yhdyssanan osat ovat erotettavissa toisistaan, kun yhdyssanan osiin merkitään sijaintitieto (ydinvoimalaitoksessa -> ydin-, -voima-, -laitos, ydinvoima-, -voimalaitos, ydinvoimalaitos). Kun osat on jo tallennusvaiheessa kytketty toisiinsa, haun tarkkuus ei kärsi, vaikka osoitteita ei olisikaan merkitty hakemistoon sanan tarkkuudella. Tällaisen yhdyssanojen käsittelyn ongelmana on, että hakemistoon tulee enemmän sekä sanoja että osoitteita kuin edellisissä vaihtoehdoissa. *n*-osaisesta yhdyssanasta syntyy hakemistoon sanoja ja osoitteita:

(5) $$\frac{n(n+1)}{2}$$

Eli kolmiosaisesta yhdyssanasta saadaan kuusi sanaa ja osoitetta, viisiosaisesta yhdyssanasta hystereesirotaatiovaihetaajuusilmaisin taas viisitoista sanaa ja osoitetta.

Hakijan kannalta merkittävämpi ongelma on, että kaikki yhdistelmät eivät tuota mielekkäitä sanoja. Esimerkiksi pystykorvarotu-sanasta syntyvä alkuosa pystykorva- on tuttu sana, mutta siitä myös muodostuva -korvarotu on omiaan vain hämmentämään tiedonhakijaa. Tämä tapa ei siis ole ymmärrettävyyden kannalta hyvä, kun käsiteltävänä on moniosaisia yhdyssanoja. Tämä vaihtoehto kuitenkin toteutettiin FULLTEXT-projektissa, jotta nähtäisiin, miten suuresti tällainen yhdyssanojen käsittely lisää hakemistosanojen ja osoitteiden kokonaismäärää.

Edelläolevissa laskelmissa siis oletettiin, että perusmuoto-ohjelma löytää käsittelemälleen yhdyssanalle vain yhden tulkinnan. Mikäli yksikin yhdys-

sanan osista puuttuu perusmuoto-ohjelman sanakirjasta, koko yhdyssana jää tulkitsematta. Tällöin yhdyssana tallennetaan taipuneessa muodossaan tunnistamattomien sananmuotojen hakemistoon. Mikäli yhdyssanaa ei entuudestaan ole tunnistamattomien sananmuotojen joukossa, hakemistoon tallentuu yksi hakemistosana ja yksi osoite.

Perusmuoto-ohjelma voi myös löytää analysoitavalle sananmuodolle useita mahdollisia tulkintoja. Suomenkielisessä tekstissä noin 15 % sananmuodoista on monitulkintaisia (Karlsson 1994, s. 80). Sananmuotojen monitulkintaisuus lisää hakemistosanojen ja osoitteiden määrää, koska kaikki tulkinnat on tallennettava hakemistoon. (Tosin ylimääräisiä tulkintoja voidaan karsia pois esimerkiksi heuristisin keinoin; tarkemmin Karlsson 1990.) Tällainen ylimääräisten muotojen tuottaminen eli **ylitulkinta** aiheuttaa myös sen, että perusmuoto-ohjelman toimintaa voi olla vaikea ymmärtää, koska ylimääräiset muodot ovat sellaisia, joita tavallinen kielenkäyttäjä ei edes tule ajatelleeksi.

Jos perusmuotohakemistosta löytyy vaikkapa paperiarkinen, kokkolapsi tai kokkolasta, hakija hakemistoa selatessaan varmaankin ihmettelee sieltä löytyviä merkillisiä perusmuotoja ja jopa taivutusmuotoja. Ensimmäinen näistä esimerkeistä on syntynyt paperiarkki-sanan elatiivimuodosta paperiarkista. Toinen ja kolmas esimerkki puolestaan ovat lähtöisin Kokkola-sanan elatiivista Kokkolasta. Perusmuoto-ohjelman tulkinnan mukaan se voi olla Kokkola-sanan taivutusmuoto, mutta myös partitiivimuoto yhdyssanasta, jonka osia ovat kokko- ja -lapsi tai nominatiivi yhdyssanasta, jonka osia taas ovat kokko- ja -lasta. Vaikka tällaiset ylitulkinnat ovat morfologisesti oikein, ne eivät ole semanttisesti mielekkäitä. Haun saannin kannalta ylitulkinta ei tuota ongelmia, koska hakemistosta löytyvät myös oikeat muodot (paperiarkki, kokkola). Sen sijaan ylimääräiset tulkinnat voivat huonontaa haun tarkkuutta (hakusanalla *-lapsi* saadaan dokumentti, jossa ei ole yhtäkään lapsi-sanan esiintymää, vaan sen sijaan sananmuotohomografi Kokkolasta).

FULLTEXT-projektin APL-Minttu-osuudessa toteutettiin hakemisto, jossa yhdyssanat palautettiin perusmuotoon ja niiden osat kukin vuorollaan ns. vyörytettiin sanan loppuun. Esimerkiksi ydinvoimalaitoksessa-sananmuodosta tallennettiin hakemistoon muodot ydinvoimalaitos ja voimalaitos,ydin sekä laitos,ydinvoima (Alkula & Honkela 1992, s. 45).

Selailun kannalta tällainen hakemisto on kätevä, kun hakija näkee jo hakemistosta, minkä sanan osana hakusana on esiintynyt. Näin kyselyyn voidaan poimia vain halutut sanat (kuten laitos,kauppaoppi) ja jättää epärelevantiksi

katsotut valitsematta (kuten laitos,hoito). Testauksissa todettiin, että tässä tapauksessa hakemisto sisältää vähemmän hakemiston merkkijonoja kuin taivutusmuotohakemisto, joten sillä on taivutusmuotohakemistoa pienempi muistitilan tarve (Alkula & Honkela 1992, s. 51). BASIS-hakujärjestelmässä vastaavaa vaihtoehtoa ei kokeiltu, koska tällaisen hakemiston kiinnostavin ominaisuus on hakusanojen ideoinnin tukeminen - hakemiston käyttökelpoisuutta ideointiin ei voi selvittää pelkin järjestelmätestein, vaan sen tutkimiseen tarvittaisiin myös todellisten tiedontarvitsijoiden tekemiä testihakuja. Tähän FULLTEXT-projektissa ei rajallisten resurssien vuoksi ollut mahdollisuuttta.

7.3 Tuotetut hakemistot

7.3.1 Toteutuksen periaatteet

FULLTEXT-projektissa käytettyjen testihakemistojen toteutusperiaatteet oli suunniteltava itse projektissa, koska tämäntyyppistä tutkimusta ei suomenkielisellä aineistolla ollut aiemmin tehty. Englanninkielistä aineistoa käyttäneitä tutkimuksiakaan ei voinut käyttää mallina, koska englanninkielisen tekstin ongelmat ovat erilaiset. Esimerkkinä tästä voidaan mainita yhdyssanat: suomenkielisen tekstin käsittelyssä ongelmana on, miten yhdyssanat pilkotaan osiinsa (ydinvoimalaitos), kun taas englanninkielisessä tekstissä tarvitaan keinot, joilla kytkeä erillisinä esiintyvät sanat yhteen yhdyssanaksi (nuclear power plant).

Edellisessä luvussa esitettyjen arviointiperusteiden avulla tutkimusta varten päätettiin toteuttaa neljä erilaista hakemistoratkaisua, joita selostetaan tarkemmin seuraavissa alaluvuissa.

Kaikki BASIS-hakujärjestelmän tutkimustietokannat tuotettiin Aamulehden kolmen kuukauden eli 89 päivän aineistosta siten, että testijärjestelmiin syötettiin kerrallaan yhden päivän aineisto. Suomen kielen morfologisia tulkintaohjelmia sovellettiin useilla eri tavoilla, mutta muuten testitietokannat tuotettiin samalla periaatteella kuin Aamulehden elektroninen arkisto, jotta testiympäristöt vastaisivat mahdollisimman hyvin oikeaa tuotantokäytössä olevaa tekstitietokantaa.

Yksi poikkeus Aamulehden tallennuskäytännöstä tehtiin: tieto hakemiston merkkijonojen sijainnista dokumentissa (eli merkkijonojen osoite) tallennettiin testihakemistoihin virkkeen tarkkuudella, kun ne Aamulehden tuotantoversiossa oli ilmaistu väljemmin eli kappaleen tarkkuudella. Tämä ratkaisu

tehtiin, jotta voitaisiin kokeilla yhdyssanojen osien kytkemistä sekä JA-operaattorilla että tarkimmalla mahdollisella läheisyysoperaattorilla, joka BASIS-ohjelmiston K-versiossa oli virkeoperaattori[1].

Tallennusvaiheessa sovellettiin Aamulehden arkistoa varten laadittua sulkusanalistaa. Jotta kaikki tutkittavat hakemistot olisivat olleet mahdollisimman yhdenmukaiset muuten paitsi tutkittavien muuttujen suhteen, Aamulehden sulkusanalistaa käytettiin jokaisen hakemiston tuottamisessa. Siten sulkusanalistalla ei ole olennaista vaikutusta hakemistojen välillä tehtyihin vertailuihin.

Tallennusvaiheessa BASIS kävi sulkusanalistan läpi ennen sananmuotojen perusmuotoistamista. Aamulehden sulkusanalistalla oli lueteltu 75 itsenäistä sanaa, lähinnä partikkeleita (kuten ja, että, jos, kun), pronomineja (minä, minun, se, joka, jonka) ja olla-verbin muotoja (olen, olin, ollut, on). Lisäksi listaan sisältyi 24 eri taivutuspäätettä (ssa, ssä, ään, iin, een). Kun ne karsittiin, hakemistoon ei joutunut irrallisia lyhenteiden tms. lopussa olleita päätteitä; BASIS-järjestelmä näet tulkitsi kaksoispisteen ja eräiden muiden erikoismerkkien erottamat sananosat erillisiksi sanoiksi (eli EY:ssä -> EY; pääte ssä ei tallennu hakemistoon).

Hakemistojen vertailu tapahtui BASIS-hakujärjestelmän tuottamien seurantatietojen avulla. Jokaisen päivitysajon eli yhden päivän aineiston syöttämisen jälkeen tulostettiin seurantatiedot, joista kävi ilmi muun muassa käsiteltyjen merkkijonojen määrä (transactions processed), uusien hakemiston merkkijonojen määrä (new terms added), uusien osoitteiden määrä (total postings added) sekä eri tiedostojen viemän muistitilan määrä.

Muistitilan ja hakemistojen tuottamiseen kuluvan ajan optimoimiseksi jaettiin BASIS-hakujärjestelmän hakemistot fyysisesti seitsemään segmenttiin. Kussakin segmentissä oli tietyillä kirjaimilla alkavat sananmuodot (esimerkiksi ensimmäisessä lohkossa a-d-kirjaimilla alkavat sanat, seuraavassa e-h-alkuiset jne.). BASIS-järjestelmän tuottamista seurantatiedoista näkyi, kuinka monta lohkoa (block) kukin segmentti vei muistitilaa päivitysajon jälkeen. (VAX-laitteiston yksi lohko varaa muistitilaa 512 merkkiä.) Laskemalla yhteen eri segmenttien kuluttama muistitilan määrä saatiin koko hakemiston tarvitseman muistitilan määrä.

[1] BASIS-järjestelmässä virkkeen rajaksi tulkitaan piste, huutomerkki tai kysymysmerkki, jota seuraa välilyöntimerkki.

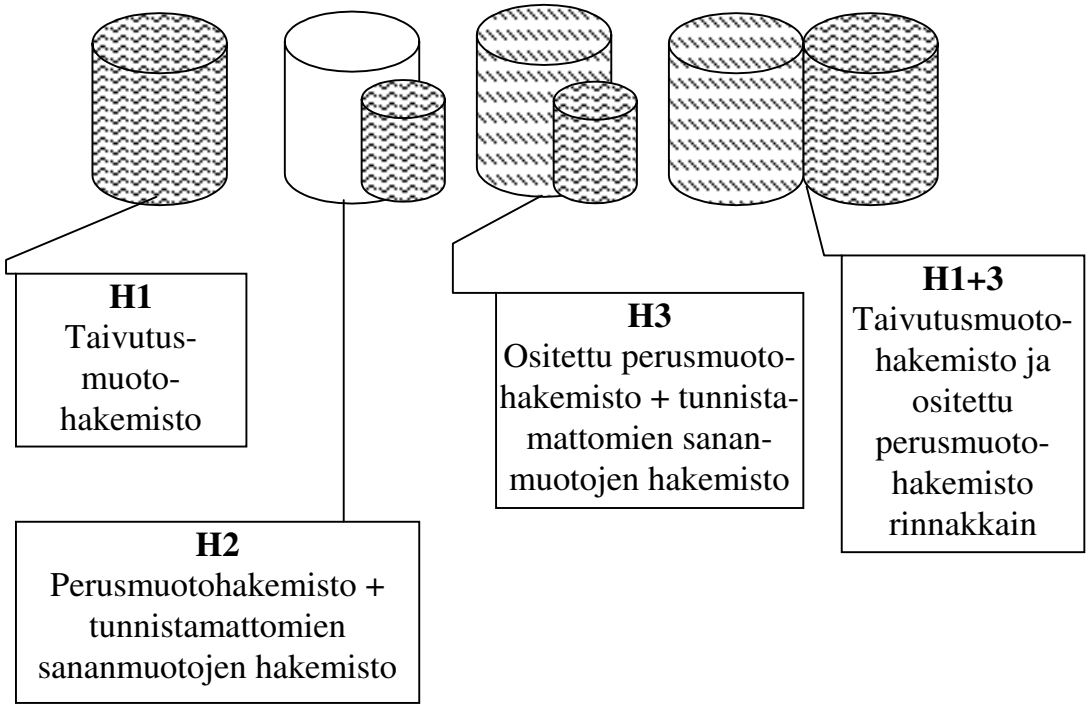
BASIS-hakujärjestelmässä erilaisia hakemistoratkaisuja oli neljä (kuva 4):

H1) Taivutusmuotohakemisto

H2) Perusmuotohakemisto täydennettynä tunnistamattomien sananmuotojen hakemistolla

H3) Ositettu perusmuotohakemisto, joka sisälsi edellisen vaihtoehdon lisäksi yhdyssanoista kaikki niiden osat sekä näiden osien yhdistelmät

H1+3) Kaksoishakemisto: ositettu perusmuotohakemisto, jonka rinnalla oli taivutusmuotohakemisto



Kuva 4. BASIS-hakujärjestelmässä toteutetut hakemistotyypit.

7.3.2 Taivutusmuotohakemisto

Ensimmäinen hakemistoista (H1) tuotettiin perinteisellä tavalla eli siinä ei käytetty minkäänlaista suomen kielen tulkintaohjelmaa. Eri kenttien sisältö tallennettiin hakemistoon sellaisenaan, tekstiä sisältäneiden kenttien sananmuodot siis taivutusmuotoisina (esimerkiksi talvisodassa).

Tämä taivutusmuotohakemisto oli perusvaihtoehto, johon toisia hakemistoja verrattiin. Sananmuotojen tallentaminen sellaisinaan hakemistoon on nykyisissä tuotantokäytössä olevissa hakujärjestelmissä sovellettu, vakiintunut

ratkaisu. On tärkeää tietää, pystytäänkö morfologisten tulkintaohjelmien avulla tuottamaan hakemistoja, joiden ominaisuudet ovat tällaista taivutusmuotohakemistoa paremmat. Jos uudenlaiset hakemistot eivät edes teoriassa ja testijärjestelmissä toimi paremmin kuin jo yleisesti käytössä oleva hakemistoratkaisu, ei ole mitään mieltä tuoda niitä tuotantokäyttöön. Jos taas uudet hakemistoratkaisut toimivat paremmin kuin nykyinen ratkaisumalli, hakujärjestelmien tuottaja joutuu pohtimaan, ovatko uusien ratkaisujen edut niin suuret, että taivutusmuotohakemisto kannattaisi korvata niillä.

7.3.3 Perusmuotohakemistot

Perusmuotohakemistoissa (H2 ja H3) syötettiin teksti- ja otsikkokenttien sisältämät sananmuodot ensin perusmuoto-ohjelmalle. Käytännössä tämä oli Morfo-ohjelma, koska Twol saatiin FULLTEXT-projektin käyttöön vasta projektin loppuvaiheessa eikä sillä ehditty tehdä täysimittaisia testiajoja koko tutkimusaineistolla. Twol-ohjelmalla tehtiin kuitenkin tiettyjä suppeita testiajoja, joilla sitä pyrittiin vertaamaan Morfon kanssa (luku 8.5).

Vain teksti- ja otsikkokenttien sisältö syötettiin Morfolle - esimerkiksi artikkelin päiväystietoja ei ole mitään mieltä syöttää perusmuoto-ohjelmalle, koska päivämäärä ei ole sellainen luonnollisen kielen sana, jota morfologiset tulkintaohjelmat osaisivat käsitellä. Morfo-ohjelman perusmuotoistamat sananmuodot tallennettiin (ositettuun) perusmuotohakemistoon. Sille tuntemattomat sananmuodot puolestaan tallennettiin tunnistamattomien sananmuotojen hakemistoon. Morfo-ohjelman käsittelemät sananmuodot siis tallennettiin joko perusmuotoistettuina (ositettuun) perusmuotohakemistoon tai sellaisinaan tunnistamattomien sanojen hakemistoon, muttei yhtaikaa molempiin.

H2-hakemistossa pelkästään perusmuotoistettiin sananmuodot ennen kuin ne tallennettiin hakemistoon (talvisodassa -> talvisota). Tällä haluttiin selvittää, miten pelkkä perusmuotoistaminen muuttaa hakemiston merkkijonoja ja siten hakemiston ominaisuuksia verrattuna taivutusmuotohakemistoon. Periaatteessa ainoan eron pitäisi olla, että perusmuotohakemistossa sanan eri taivutusmuotoja edustaa vain niiden yhteinen perusmuoto. Käytännössä asia ei kuitenkaan ole näin yksinkertainen, vaan homografit ja muut ongelmatapaukset vaikuttavat osaltaan siihen, millaisia hakemistosanoja perusmuotohakemistoon päätyy.

Ositetussa perusmuotohakemistossa H3 perusmuotoistamisen lisäksi hajoitettiin yhdyssanat osiinsa ja nämä osat sekä osien yhdistelmät tallennettiin

hakemistoon. Yhdyssanan osiin liitettiin tunnus, joka osoitti, millä kohtaa yhdyssanaa kyseinen sana oli esiintynyt (talvisodassa -> talvisota, talvi-, -sota). Näin haluttiin selvittää, miten yhdyssanan osat vaikuttavat hakutuloksen saantiin ja tarkkuuteen, kun ne lisätään kyselyyn. Virhetulkintojen korjaamisen helpottamiseksi haluttiin, että ositettu perusmuotohakemisto sisältää myös kokonaiset yhdyssanat (vrt. luku 10.3).

7.3.4 Kaksoishakemisto

Viimeinen hakemisto (H1+3) oli yhdistelmä hakemistoista H1 ja H3. Ositettua perusmuotohakemistoa H3 täydentävää tunnistamattomien sananmuotojen hakemistoa ei ollut lainkaan, vaan sen sijasta koko taivutusmuotohakemisto H1. Tämä sisälsi kaikki tekstissä esiintyneet sananmuodot sellaisinaan. Nämä kaksi hakemistoa siis olivat päällekkäisiä eivätkä toisiaan täydentäviä, kuten edellisessä luvussa kuvatuissa perusmuotohakemistoissa. Itse asiassa H2- ja H3-vaihtoehdoissa käytetty tunnistamattomien sananmuotojen hakemisto on taivutusmuotohakemiston osajoukko.

Kaksoishakemiston ideana on, että jos kysely perusmuotohakemistosta ei onnistu, suoritetaan uusi kysely (ts. kysely uudelleen, mutta muokattuna) taivutusmuotohakemistosta. Korjauskyselyssä voidaan perinteisellä tavalla käyttää hakijan katkaisemia hakusanoja - edellyttäen, että hakija osaa käyttää perinteistä tiedonhakujärjestelmää ja hakea taivutusmuotohakemistosta - tai hakusanat voidaan katkaista automaattisesti vartalo-ohjelmien avulla. Kahden hakemiston tuottaminen kuitenkin vaatii paljon muistitilaa. Tämän muistitilan määrä haluttiin FULLTEXT-projektissa selvittää.

H1+3 -kaksoishakemiston tuottamistapa poikkesi muiden hakemistojen tuottamisesta. Hakemistot H1, H2 ja H3 tuotettiin tuotantokäyttöä vastaavalla tavalla eli siten, että kunkin päivän artikkelit lisättiin tietokantaan omana ryhmänään. Näin voitiin vaihe vaiheelta seurata, miten hakemistot muodostuvat kolmen kuukauden ajanjaksolla. Kaksoishakemisto tehtiin niin, että ensin kopioitiin Morfo-ohjelman avulla tuotettu H3-hakemisto eli ositettu perusmuotohakemisto sellaisenaan. Sitten BASIS-järjestelmä kävi dokumenttitiedoston teksti- ja otsikkokentät uudestaan läpi ilman Morfo-ohjelmaa ja tuotti näistä sananmuodoista taivutusmuotohakemiston, jolla korvattiin varsinaisen H3-hakemiston tunnistamattomien sananmuotojen hakemisto.

Koska kaksoishakemistoa ei tuotettu samalla tavalla kuin muita hakemistoja, sen kasvua ei verrata toisten hakemistojen kasvuun (luku 8). Sen sijaan

valmiin H1+3-hakemiston tilantarvetta sekä hakuominaisuuksia verrattiin toisten hakemistojen vastaavien ominaisuuksien kanssa. Koska kaksoishakemisto tuotettiin täydentämällä H3-hakemistoa eikä nollatilanteesta kuten toisia hakemistoja, sen muistitilan käyttöäkään ei voitu optimoida yhtä hyvin kuin toisissa hakemistoissa, mikä myös on otettava huomioon tulosten tulkinnassa.

7.4 Tutkimusympäristöt

7.4.1 Toteutuksen periaatteet

Projektissa tutkimusympäristöjä oli enemmän kuin hakemistoja, koska taivutusmuotohakemistossa kokeiltiin useita erilaisia tapoja hyödyntää suomen kielen morfologisia tulkintaohjelmia.

Tutkimusympäristöt olivat seuraavat (kuva 5):

T1) Perinteinen hakeminen: hakijan itse katkaisemat hakusanat
(kysely kohdistui taivutusmuotohakemistoon H1)

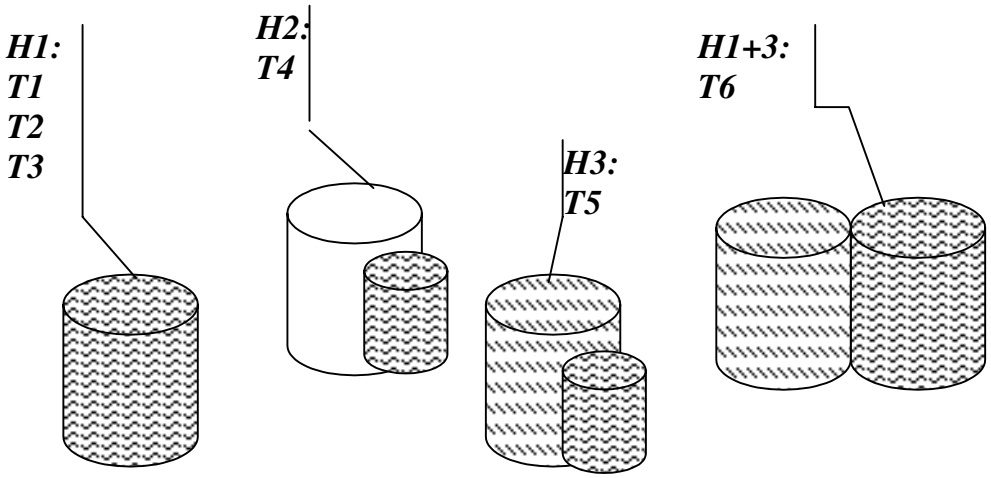
T2) Automaattinen katkaisu: perusmuotoisten hakusanojen syöttäminen taivutusvartalo-ohjelmille ja näiden tuottamien vartaloiden käyttö hakusanoina
(kysely kohdistui taivutusmuotohakemistoon H1)

T3) Seulonta: perusmuotoisten hakusanojen syöttö taivutusvartaloohjelmille, vartaloihin täsmäävien hakemiston sananmuotojen poiminta, näiden sananmuotojen tarkistaminen perusmuotoohjelmilla ja lopulta kysely seulan läpäisseillä hakemiston sananmuodoilla
(kysely kohdistui taivutusmuotohakemistoon H1)

T4) Perusmuotojen ja yhdyssanojen alkuosien hakeminen
(kysely kohdistui perusmuotohakemistoon H2)

T5) Perusmuotojen ja yhdyssanan kaikkien osien hakeminen
(kysely kohdistui ositettuun perusmuotohakemistoon H3)

T6) Perusmuotojen ja yhdyssanan osien hakeminen - jos perusmuodoilla suoritetun kyselyn tuloksena oli tyhjä joukko, hakusanojen automaattinen katkaisu ja hakeminen taivutusmuotohakemistosta
(kysely kohdistui kaksoishakemistoon H1+3)



Kuva 5. BASIS-hakujärjestelmässä toteutetut tutkimusympäristöt

Ensimmäinen vaihtoehto, T1, oli hakea hakijan katkaisemilla hakusanoilla taivutusmuotohakemistosta (katkaisuperiaatteet on kuvattu luvussa 7.6). Se oli perusvaihtoehto, johon toisia hakutapoja verrattiin. Tämä oli tärkeää, jotta saataisiin tietää, miten uudet hakumenetelmät toimivat verrattuna jo tuotantokäytössä olevaan, yleisesti käytettyyn hakumenetelmään.

Seuraavassa tutkimusympäristössä T2 tutkittiin hakusanojen automaattista katkaisua suomen kielen vartalo-ohjelmien avulla. Tämän vaihtoehdon toimintaperiaate vastaa anglo-amerikkalaisissa tiedonhakututkimuksissa käytettyjen karsinta-algoritmien toimintaperiaatteita (luku 4.3.2): hakija antaa vain hakusanan perusmuodon ja karsinta-algoritmi huolehtii hakusanan taivutusmuotojen (usein myös johdosten) hakemisesta. Siksi FULLTEXT-projektissa haluttiin kokeilla, miten hakusanojen automaattinen katkaisu toimii englannista poikkeavassa kielessä. Tosin englannin kieltä varten kehitetyt algoritmit tavallisesti poistavat sanoista sekä johtimia että taivutuspäätteitä, kun taas suomen kielen taivutusvartaloita tuottavat ohjelmat poistavat vain taivutuspäätteet (tai osan päätteistä).

Kolmantena vaihtoehtona oli seulonta (T3). Vaikka hakusanojen automaattinen katkaisu periaatteessa parantaa hakutulosten tarkkuutta, koska vartalo-ohjelmien tuottamat vartalot ovat yleensä pitempiä kuin hakijan itse katkaisemat hakusanat, myös vartalo-ohjelmien katkaisemat hakusanat tuottavat tulosjoukkoon epärelevantteja dokumentteja. Tämä johtuu siitä, että hakemiston merkkijono voi olla jonkin muun kuin itse hakusanan esiintymä, vaikka se sattuisikin alkamaan samalla merkkijonolla kuin hakusana.

Seulonnan tarkoituksena on ratkaista tämä tarkkuusongelma niin, että ennen varsinaista kyselyä tutkitaan, onko katkaistun hakusanan (eli esiintymätason

merkkijonokaavion) kanssa täsmäävä hakemistosana todellakin ilmaisutason hakusanan esiintymä vai ei. Vartalo-ohjelman tuottamia katkaistuja hakusanoja täsmäytetään ensin hakemiston merkkijonoihin. Täsmäävät merkkijonot otetaan tarkempaan käsittelyyn eli syötetään perusmuoto-ohjelmalle. Jos perusmuoto-ohjelman tuottama perusmuoto on sama kuin hakijan antama alkuperäinen perusmuoto, hakemiston merkkijono läpäisee seulan.

Neljäntenä vaihtoehtona oli hakeminen perusmuodoilla perusmuotohakemistosta. T4-ympäristöä voidaan pitää myös eräänlaisena perusvaihtoehtona T1-ympäristön rinnalla, koska se on perusmuoto-ohjelmien soveltamisen minimivaihtoehto: hakemistossa on lukuisten taivutusmuotojen sijasta niiden yhteinen perusmuoto, mutta muuta morfologista käsittelyä dokumenteissa esiintyville sananmuodoille ei ole tehty. T4-ympäristössä voidaan siis tutkia, mikä vaikutus pelkästään taivutusmuotojen normalisoinnilla on tiedonhakuun.

Viidentenä vaihtoehtona oli hakeminen ositetusta perusmuotohakemistosta (T5). Tässä vaihtoehdossa hyödynnettiin perusmuoto-ohjelmien kykyä jakaa yhdyssanat osiinsa.

Viimeinen vaihtoehto (T6) eli hakeminen kaksoishakemistosta otettiin tutkimukseen mukaan siksi, että hakemistoon tallennettavien sananmuotojen perusmuotoistamisessa saattaa tapahtua virheitä. Ongelman ratkaisuksi on ehdotettu, että hakujärjestelmässä on kaksi hakemistoa, perusmuotohakemisto ja taivutusmuotohakemisto (Hjorth 1987b). Mikäli perusmuodoilla hakeminen ei onnistu esimerkiksi siksi, että perusmuoto-ohjelma on tulkinnut homografisen sananmuodon väärin, suoritetaan kysely uudelleen hakijan katkaisemilla hakusanoilla taivutusmuotohakemistosta, johon kaikki sananmuodot on tallennettu juuri siinä muodossa kuin ne tekstissä esiintyivät.

7.4.2 Perinteinen hakeminen

Tutkimusympäristössä T1 haut tehtiin perinteiseen tapaan eli hakusanat katkaistiin itse. Kyselyt laadittiin tavalla, joka vastaa hyvää ammattikäytäntöä. Tässä tutkijan apuna oli kokeneiden informaatikkojen raati. (Hakusanojen valinta- ja katkaisuperiaatteet on selostettu tarkemmin luvussa 7.6.)

T1-ympäristössä saadut tulokset olivat perusjoukko, johon muissa tutkimusympäristössä saatuja tuloksia suhteutettiin. Kaikissa muissa tutkimusympäristöissä sovellettiin suomen kielen morfologisia tulkintaohjelmia ainakin hakuvaiheessa eli niissä käytettiin perusmuodossa olevia hakusanoja.

7.4.3 Automaattinen katkaisu

T2-tutkimusympäristössä kokeiltiin hakusanan automaattista katkaisua. Tätä varten laadittiin aliohjelma, jonka syötteenä oli perusmuotoinen hakusana. Aliohjelma lähetti hakusanan taivutusvartaloita tuottavalle ohjelmalle ja sitten lisäsi vartalo-ohjelman tuottamat vartalot kyselyyn alkuperäisen hakusanan rinnalle. Tämä vaihtoehto toteutettiin sekä Finstems- että Hahmotin-ohjelmilla (kuvat 6 ja 7).

Aluksi ilmaisutason hakusuunnitelmassa olleet hakusanat syötettiin yksi kerrallaan aliohjelmalle. Syötteenä olivat hakusanan perusmuoto sekä sen sanaluokkakoodi, jotka kirjoitettiin vartalo-ohjelman edellyttämällä tavalla. Esimerkiksi hakusana *autoverotus* syötettiin Finstems-ohjelmaa käytettäessä muodossa n:auto/verotus.

Vartalo-ohjelma tuotti joukon taivutusvartaloita, *autoverotus*-hakusanan tapauksessa siis vartalot autoverotus, autoverotuks, autoverotude ja autoverotut. Näiden loppuun aliohjelma liitti BASIS-hakujärjestelmän yleiskatkaisu-

```
stem
Give the Input (eg. V:istua):
n:kauppa

  60/
Hae...
*       1147   60/  TEKSTI=kauppa* (  348   TERMIÄ YHDISTETTY)
*        100   61/  kaupoi* (  6   TERMIÄ YHDISTETTY)
*         34   62/  kauppoi* (  5   TERMIÄ YHDISTETTY)
*        120   63/  kauppoj* (  3   TERMIÄ YHDISTETTY)
*        907   64/  kaupa* (  112   TERMIÄ YHDISTETTY)
*       1872   65/  TEKSTI=kauppa* TAI  kaupoi* TAI  kauppoi* TAI kauppoj+
```

Kuva 6. Kysely Finstems-ohjelman avulla. STEM-komento kutsuu aliohjelmaa, joka käynnistää Finstems-ohjelman suorituksen. Tulostuksen toinen sarake kertoo, montako dokumenttia kullakin hakuavaimella saatiin yhteensä. Seuraava sarake kertoo tulosjoukon numeron (korin numero), sen jälkeen on itse hakusana. Etuliite TEKSTI= kertoo, että kysely kohdistui tekstikenttään. Rivin lopussa suluissa kerrotaan, kuinka monen eri hakemistosanan osoitetiedot tulosjoukkoon on yhdistetty (esimerkiksi kori numero 60 on muodostunut 348:sta eri hakemistosanasta, jotka täsmäävät katkaistun kauppa-hakusanan kanssa). Alimmalla rivillä näkyy kaikkien eri hakusanoilla saatujen korien unioni (+-merkki rivin lopussa tarkoittaa, että hakusanoja on enemmän kuin yhdelle riville on mahtunut näkyviin).*

merkin eli tähden *****. Sen jälkeen aliohjelma haki taivutusmuotohakemistosta näihin vartaloihin täsmäävät hakemiston merkkijonot kyselyllä:

> HAE *autoverotus** TAI *autoverotuks** TAI *autoverotude** TAI *autoverotut**

Taivutusvartaloita tuottavan ohjelman liittäminen nykyisiin, taivutusmuotohakemistoja sisältäviin tiedonhakujärjestelmiin on teknisesti varsin yksinkertaista. Se ei edellytä muutoksia hakemistoihin, vaan on hakuvaiheessa valittavissa oleva lisäominaisuus.

7.4.4 Seulonta

T3-tutkimusympäristössä tutkittiin, voidaanko hakutuloksen tarkkuutta parantaa automaattisesti niin, että kyselyn tuloksena saaduista dokumenteista

```
hahmo
HAHMOTIN - Hakuvartaloiden tuotin
Esimerkkisanoja :
    PULMA
    v TAITAA
 cc    a TARKKA
...kauppa

  23/
Hae...
*     907    23/  TEKSTI=kaupa* (  112   TERMIÄ YHDISTETTY)
*     1147   24/  kauppa* (  348  TERMIÄ YHDISTETTY)
*      149   25/  kauppo*  (  8  TERMIÄ YHDISTETTY)
*      100   26/  kaupoi* (  6  TERMIÄ YHDISTETTY)
*     1872   27/  TEKSTI=kaupa* TAI  kauppa* TAI kauppo* TAI kau+
```

Kuva 7. Kysely Hahmotin-ohjelman avulla. HAHMO-komento kutsuu aliohjelmaa, joka käynnistää Hahmotin-ohjelman suorituksen. Tulostuksen toinen sarake kertoo, montako dokumenttia kullakin hakuavaimella saatiin yhteensä. Seuraava sarake kertoo tulosjoukon numeron (korin numero), sen jälkeen on itse hakusana. Etuliite TEKSTI= kertoo, että kysely kohdistui tekstikenttään. Rivin lopussa suluissa kerrotaan, kuinka monen eri hakemistosanan osoitetiedot tulosjoukkoon on yhdistetty (esimerkiksi kori numero 23 on muodostunut 112:sta eri hakemistosanasta, jotka täsmäävät katkaistun kaupa-hakusanan kanssa). Alimmalla rivillä on kaikkien eri hakusanoilla saatujen korien unioni (+-merkki rivin lopussa tarkoittaa, että hakusanoja on enemmän kuin yhdelle riville on mahtunut näkyviin).*

hyväksytään vain sellaiset, joissa aidosti esiintyy jokin hakusanan muoto. Muut karsitaan pois: siis sellaiset dokumentit, joissa esiintyy jonkin muun sanan muoto, mutta joka sattuu alkamaan samalla merkkijonolla kuin hakusana.

Seulontaa varten rakennettiin aliohjelma, joka ensin kysyi hakijalta hakusanan perusmuodon ja sanaluokan. Nämä (esimerkiksi n:tarkastus) syötettiin Finstems-ohjelmalle, joka tuotti niistä vartalot. Tähän asti siis edettiin kuten edellisessä T2-tutkimusympäristössä.

Seuraavaksi katsottiin, mihin hakemiston merkkijonoihin Finstemsin tuottamat vartalot täsmäsivät. Tässä hyödynnettiin BASIS-hakujärjestelmän KATSO-komentoa, jolla voidaan selata hakemiston merkkijonoja ilman varsinaista kyselyn suorittamista (KATSO-komento ei siis muodosta tulosjoukkoja). Esimerkiksi *tarkastus*-hakusanasta syntyivät vartalo-ohjelman avulla seuraavat katkaistut hakusanat: *tarkastus**, *tarkastuks**, *tarkastude** ja *tarkastut**. Näillä poimittiin hakemistosta muun muassa seuraavat hakemiston merkkijonot: tarkastus, tarkastusalukset, tarkastusasemalla, tarkastushetkellä, tarkastusilmoitusta … tarkastukseen, tarkastuksen, tarkastuksessa jne.

Seuraavaksi aliohjelma syötti kunkin hakemistosta saadun sananmuodon Morfo-ohjelmalle, joka tuotti siitä perusmuodon. Tämän jälkeen Morfon tuottamaa perusmuotoa ja alkuperäistä, perusmuodossa annettua hakusanaa (ilmaisutason hakusanaa) verrattiin keskenään. Mikäli ne olivat täsmälleen samat, taivutusmuotohakemistosta saatu sananmuoto kelpuutettiin jatkoon. Mikäli taas hakemistosanan perusmuoto ei täsmännyt alkuperäisen hakusanan kanssa tai se jäi kokonaan tunnistamatta, hakemistosta saatu sananmuoto hylättiin.

Kun edellä kuvatut tarkistukset oli tehty eli seulonta suoritettu, tehtiin varsinainen kysely eli haettiin myös dokumentit. Kyselyssä annettu alkuperäinen hakusana korvattiin seulonnan läpäisseillä hakemiston merkkijonoilla, jotka kytkettiin toisiinsa TAI-operaattorilla. Näitä taivutusmuotoisia hakemiston merkkijonoja ei siis enää tässä vaiheessa katkaistu, vaan kyselyssä käytettiin täsmälleen niitä hakusanan esiintymiä, jotka jo tiedettiin hakemistossa olevan (Kuva 8). Mikäli hakusuunnitelmassa oli useita hakusanoja, jokainen seulottiin erikseen edellä kuvatulla tavalla.

```
23/ /sanat
Give the Input (eg. V:istua):
n:leiri/koulu
leirikoulu*
 nument = 11
 more   = 0
TEKSTI=LEIRIKOULUASIOISTA
TEKSTI=LEIRIKOULUIHIN
TEKSTI=LEIRIKOULUISTA
TEKSTI=LEIRIKOULUJA
TEKSTI=LEIRIKOULUJEN
TEKSTI=LEIRIKOULUJUTUISSA
TEKSTI=LEIRIKOULUMATKALLA
TEKSTI=LEIRIKOULUN
TEKSTI=LEIRIKOULUSSA
TEKSTI=LEIRIKOULUSTAAN
TEKSTI=LEIRIKOULUUN
  23/
Hae...
*       1   23/  TEKSTI=LEIRIKOULUASIOISTA
*       1   24/  TEKSTI=LEIRIKOULUIHIN
*       1   25/  TEKSTI=LEIRIKOULUISTA
*       1   26/  TEKSTI=LEIRIKOULUJA
*       1   27/  TEKSTI=LEIRIKOULUJEN
*       1   28/  TEKSTI=LEIRIKOULUJUTUISSA
*       1   29/  TEKSTI=LEIRIKOULUMATKALLA
*       2   30/  TEKSTI=LEIRIKOULUN
*       1   31/  TEKSTI=LEIRIKOULUSSA
*       1   32/  TEKSTI=LEIRIKOULUSTAAN
*       1   33/  TEKSTI=LEIRIKOULUUN
*    10   34/  TEKSTI=LEIRIKOULUASIOISTA TAI TEKSTI=LEIRIKOUL+
```

Kuva 8. Kysely seulotuilla sananmuodoilla. SANAT-komennon käynnistämä aliohjelma katkaisi hakusanan Finstems-ohjelmalla, jonka jälkeen hakemistosta poimittiin näin saatuihin vartaloihin täsmäävät merkkijonot. Ylemmässä listauksessa ovat nämä täsmäävät hakemiston merkkijonot. Kukin hakemiston merkkijono syötettiin Morfo-ohjelmalle, jonka tuottamaa perusmuotoa verrattiin alkuperäiseen hakusanaan. Jos ne olivat samat, hakemiston merkkijono hyväksyttiin ja lisättiin kyselyyn.- Sen jälkeen suoritettiin itse kysely (HAE-komento). Alemmassa listauksessa näkyvät kunkin hakusanan osumat. Listauksen toinen sarake kertoo, montako dokumenttia kullakin hakusanalla löydettiin yhteensä. Seuraavassa sarakkeessa on tulosjoukon numero (korin numero), sitten itse hakusana (etuliite TEKSTI= kertoo, että kysely kohdistui tekstikenttään). Alimmalla rivillä on kaikkien eri hakusanoilla saatujen korien unioni (+-merkki rivin lopussa tarkoittaa, että hakusanoja on enemmän kuin yhdelle riville on mahtunut näkyviin).

Edellä kuvatun seulonnan tyyppistä ideaa on sovellettu esimerkiksi CITE/ CATLINE-näyttöluettelossa. Siinä tosin ei verrattu perusmuotoja vaan vartaloita. Ensin hakusanoista karsittiin pois päätteet, jonka jälkeen hakusuunnitelman hakusanat korvattiin kyselyssä karsinta-algoritmin tuottamilla vartaloilla. Katkaisu kuitenkin huononsi tulosjoukon tarkkuutta, koska katkaistut vartalot täsmäsivät muihinkin sanoihin kuin varsinaisen hakusanan esiintymiin. Niinpä hakemistosta löytyneet sananmuodot syötettiin edelleen katkaisualgoritmin käsiteltäväksi. Jos se tuotti hakemistosanasta saman vartalon kuin alkuperäisestäkin hakusanasta, hakemistosana hyväksyttiin, muussa tapauksessa hylättiin. (Ulmschneider & Doszkocs 1983)

7.4.5 Perusmuotojen ja yhdyssanojen alkuosien hakeminen

Neljännessä ympäristössä (T4) kysely kohdistui perusmuotohakemistoon (H2). Hakusanat syötettiin perusmuodossa hakujärjestelmälle, joka poimi hakemistosta niiden kanssa täsmäävät hakemistosanojen perusmuodot.

Kun haluttiin myös hakusanalla alkavat yhdyssanat, apuna käytettiin Finstems-ohjelmaa. Ensin perusmuotoinen hakusana ja sen sanaluokka syötettiin Finstems-ohjelmalle, joka tuotti hakusanasta taivutusvartalot. Taivutusvartaloiden loppuun liitettiin automaattisesti katkaisumerkki *, minkä jälkeen katkaistut hakusanat haettiin sekä perusmuotohakemistosta että sitä täydentävästä tunnistamattomien sananmuotojen hakemistosta (kuva 9). Yhdyssanojen hakeminen siis toteutettiin samalla tekniikalla kuin hakusanojen automaattinen katkaisu tutkimusympäristössä T2.

Yhdyssanat haettiin myös tunnistamattomien sananmuotojen hakemistosta, jotta kysely kattaisi mahdollisimman monet hakusanan esiintymät. Tällä tavalla löydetään nekin yhdyssanat, joita Morfo ei jostain syystä ole tallennusvaiheessa tunnistanut; esimerkiksi, jos hakusana on esiintynyt sellaisen yhdyssanan alussa, jonka loppuosana ollutta sanaa ei ollut Morfo-ohjelman sanakirjassa.

7.4.6 Perusmuotojen ja yhdyssanojen kaikkien osien hakeminen

T5-tutkimusympäristössä hakusanat syötettiin perusmuodossa hakujärjestelmälle, joka poimi hakemistosta niiden kanssa täsmäävät hakemistosanojen perusmuodot.

```
stem
Give the Input (eg. V:istua):
n:maa/talous
/
Hae...
*     417    24/  TEKSTI=maatalous* (  142   TERMIÄ YHDISTETTY)
*      0    25/ maatalouks*
*      3    26/ maataloude* (  2  TERMIÄ YHDISTETTY)
*      0    27/ maatalout*
*     419   28/  TEKSTI=(maatalous* TAI maatalouks* TAI  maatalou+
/
stemzt
Give the Input (eg. V:istua):
n:maa/talous
/
Hae...
*      4    29/  ZT=maatalous* (  4   TERMIÄ YHDISTETTY)
*      0    30/ ZT=maatalouks*
*      0    31/ ZT=maataloude*
*      0    32/ ZT=maatalout*
*      4    33/ ZT=(maatalous* TAI maatalouks* TAI  maataloude* TAI+
/
34/ HAE 28 TAI 33
*      419    34/  28 tai 33
```

Kuva 9. Yhdyssanojen hakeminen T4-ympäristössä. STEM-komennolla käynnistetään aliohjelma, joka katkaisee hakusanat Finstems-ohjelman avulla ja hakee sitten näillä vartaloilla alkavat sanat perusmuotohakemistosta. STEMZT-komennolla käynnistetään aliohjelma, joka hakee vastaavilla vartaloilla tunnistamattomien sananmuotojen hakemistosta. - Toinen sarake kertoo, montako dokumenttia kullakin hakuavaimella saatiin. Seuraava sarake kertoo tulosjoukon numeron (korin numero), sen jälkeen on itse hakusana. Rivin lopussa suluissa kerrotaan, kuinka monen eri hakemistosanan osoitetiedot tulosjoukkoon on yhdistetty (esimerkiksi kori numero 26 on muodostunut kahdesta eri hakemistosanasta, jotka täsmäävät katkaistun maatalous-hakusanan kanssa). Etuliite TEKSTI= ilmaisee, että kysely kohdistui perusmuotohakemiston tekstikenttään, etuliite ZT= taas että kysely kohdistui taivutusmuotohakemiston tekstikenttään. Näiden jälkeen on eri hakusanoilla hakemistosta saatujen korien unioni (+-merkki rivin lopussa tarkoittaa, että hakusanoja on enemmän kuin yhdelle riville on mahtunut näkyviin). Toiseksi alimmalla rivillä on manuaalisesti muodostettu näistä kahdesta hakemistosta saatujen tulosjoukkojen unioni.*

```
54/ hae teksti=maatalous
*     187    54/  TEKSTI=maatalous
55/ hae teksti=maatalous-
*     325    55/  TEKSTI=maatalous-
56/ hae teksti=-maatalous-
*      5      56/  TEKSTI=-maatalous-
57/ hae teksti=-maatalous
*      4     57/  TEKSTI=-maatalous
58/ hae 54 TAI 55 TAI 56 TAI 57
*     418    58/  54 TAI 55 TAI 56 TAI 57
```

Kuva 10. Yhdyssanojen hakeminen ositetusta perusmuotohakemistosta (T5-ympäristö). Yhdyssanojen katkaisukohtaa symboloivat merkit on lisätty hakusanoihin manuaalisesti. Tähdellä alkavan rivin toinen sarake kertoo, montako dokumenttia kullakin hakuavaimella saatiin. Seuraava sarake kertoo tulosjoukon numeron (korin numero), sen jälkeen on itse hakusana. Alimmalla rivillä on muodostettu kaikkien eri hakusanoilla saatujen korien unioni.

Yhdyssanoja sisältävät dokumentit haettiin siten, että hakusanaan lisättiin manuaalisesti yhdyssanan eri osia symboloiva katkaisumerkki, jollaiseksi oli määritelty tavuviiva (*auto* -> *auto*-, *-auto*-, *-auto*). Nämä täsmäsivät hakemistosanoihin, joihin oli liitetty vastaavat yhdyssanan katkaisukohtaa ilmaisevat merkit. Lopuksi perusmuodolla ja yhdyssanan eri osilla saaduista tulosjoukoista muodostettiin manuaalisesti unioni yhdistämällä tulosjoukot yhdeksi tulosjoukoksi TAI-operaattorin avulla (kuva 10).

Edellisessä luvussa kuvatussa T4-ympäristössä yhdyssanojen alkuosat haettiin katkaistuilla hakusanoilla myös tunnistamattomien sananmuotojen hakemistosta. T5-testausympäristössä ei tehty vastaavaa yhdyssanojen hakemista tunnistamattomien sanojen hakemistosta. Tarkoituksena oli vertailla sen ja T4-ympäristön tuloksia, jotta voitaisiin nähdä, miten hyvin pelkästään perusmuotohakemistosta löydetään kaikki halutun yhdyssanan sisältävät dokumentit.

Lisäksi T4- ja T5-tutkimusympäristöissä sovellettiin ongelmakyselyissä virheenkorjausmenetelmiä, joilla pyrittiin korjaamaan tallennusvaiheessa tehdyt väärät tulkinnat, ts. poimimaan väärin perusmuotoistetut sananmuodot esiin perusmuotohakemistosta (näistä tarkemmin luvussa 7.5).

7.4.7 Hakeminen kaksoishakemistosta

Viimeisessä tutkimusympäristössä (T6) ei vakiokyselyjä enää suoritettu, koska tulokset olisivat olleet täsmälleen samat kuin T5-ympäristössä. Sen sijaan ongelmakyselyjen käsittelyssä tutkimusympäristö T6 poikkesi T5-ympäristöstä.

Kun perusmuotohakemistosta (T4) ja ositetussa perusmuotohakemistosta haettaessa (T5) törmättiin hakusanaan, joka puuttui perusmuoto-ohjelman sanakirjasta, sovellettiin virheenkorjausmenetelmiä, joilla väärin tulkitut sananmuodot pyrittiin jäljittämään perusmuotohakemistoista (luku 7.5). Sen sijaan kaksoishakemistoon perustuvassa T6-ympäristössä toimittiin toisin: siinä vain yksinkertaisesti haettiin ongelmasana taivutusmuotohakemistosta, jossa sananmuodot olivat siinä muodossa kuin ne olivat alkuperäistekstissäkin.

T6-ympäristön korjauskyselyt voidaan toteuttaa eri tavoin. Suoraviivaisin tapa on tehdä kysely perinteiseen tapaan hakijan katkaisemilla hakusanoilla (kuten T1-tutkimusympäristössä). Tämä ei kuitenkaan ole toimiva ratkaisu tapauksissa, joissa hakijat eivät ole tottuneet hakemaan taivutusmuotohakemistosta. Tällöin hakusanat on katkaistava automaattisesti (kuten T2:ssa) ja mahdollisesti vielä seulottava tulokset (kuten T3-ympäristössä).

Koska FULLTEXT-projektissa oli tarkoituksena tutkia perusmuotoisia hakusanoja, myös T6-ympäristön korjauskyselyt haluttiin toteuttaa perusmuotoisilla hakusanoilla. Siten toimintatavaksi valittiin perusmuodossa annettujen hakusanojen automaattinen katkaisu eli T2-ympäristön menetelmä.

7.5 Syötteen tarkistus- ja virheenkorjausmenetelmien yleisperiaatteet

Perinteisellä tavalla toteutetussa hakujärjestelmässä syötteen oikeellisuuden varmistaminen on täysin käyttäjän vastuulla. Hakujärjestelmä tosin antaa virheilmoituksen, mikäli kyselyn syntaksissa on korjaamista (esimerkiksi operaattoria on käytetty virheellisesti), mutta se ei pysty arvaamaan, onko hakusana katkaistu oikein. Tämä on käyttäjän itsensä pääteltävä saamiensa tulosten perusteella. Kokemuksen kautta perinteisen hakujärjestelmän käyttäjä oppii peukalosääntöjä, miten virhetilanteissa pitäisi menetellä – esimerkiksi tarkistamaan, onko hakusanassa kirjoitusvirhe, jos kyselyn tuloksena saatu tulosjoukko on tyhjä.

Suomen kielen morfologisilla tulkintaohjelmilla voidaan tarkistaa hakusanojen oikeellisuus. Käytännössä tarkistusmenetelmät riippuvat sekä hakujärjestelmän ominaisuuksista että tulkintaohjelmien soveltamistavasta. Esimerkiksi BASIS-K-hakujärjestelmän toteutuksesta näkyi, että sen tuottajat olivat suunnitelleet ohjelmiston englannin kielen mukaan. Vaikka BASIS-K oli modulaarisesti rakennettu ohjelmisto, johon oli periaatteessa mahdollista liittää muita ohjelmia, voitiin morfologisia tulkintaohjelmia käytännössä käyttää vain tietyissä tallennus- tai hakuprosessin vaiheissa, jotka eivät suomen kielen kannalta olleet parhaita mahdollisia.

Korjauskyselyissä BASIS-hakujärjestelmälle syötetyt hakusanat tarkistettiin ja muokattiin kussakin tutkimusympäristössä käyttökelpoiseen muotoon FULLTEXT-projektia varten laadittujen aliohjelmien avulla. Korjauskyselyissä oletettiin, että hakija tuntee tulkintaohjelmien toimintaperiaatteet ja osaa antaa hakusanat ja niiden sanaluokat oikein.

Aliohjelmat keräsivät yhden morfologisen tulkintaohjelman tulokset, tarvittaessa välittivät ne edelleen seuraavalle morfologiselle tulkintaohjelmalle, ja lopulta toteuttivat kyselyn tulkintaohjelmien tuottamilla hakusanoilla. Siirtymiä hakuohjelman ja tarkistuksia tekevien suomen kielen morfologisten tulkintaohjelmien välillä ei tutkimusympäristöissä kuitenkaan toteutettu niin viimeistellysti kuin tuotantokäytössä olevissa hakujärjestelmissä pitäisi tehdä. Esimerkiksi kyselyjen kaikkia vaiheita ei aina suoritettu automaattisesti, jos tarvittavan aliohjelman laatiminen olisi vaatinut suhteettoman paljon resursseja. Tuotantojärjestelmissä tällainen kyselyjen muodostaminen ja tulosjoukkojen yhdistäminen pitäisi tehdä automaattisesti.

Tutkimuksessa kokeillut korjausperiaatteet ovat toki sovellettavissa myös tuotantokäytössä olevissa hakujärjestelmissä. Morfologiset tulkintaohjelmat vain pitää integroida hakujärjestelmään paremmin ja käyttöliittymät laatia todellista käyttäjää eikä tutkijaa varten.

Kun kysely kohdistuu perusmuotohakemistoon, on päätettävä, missä vaiheessa hakuprosessia hakusanan oikeellisuus tarkastetaan. Vaihtoehtoja on kaksi:

a) Tutkitaan ensimmäiseksi perusmuoto-ohjelman sanakirjasta, löytyykö hakusana sellaisenaan sieltä. Jos hakusana löytyy suoraan sanakirjasta, suoritetaan sillä kysely perusmuotohakemistosta. Jos taas ei löydy yhtä yksiselitteistä perusmuototulkintaa, tehdään lisätarkistukset.

b) Haetaan hakijan antama hakusana ensin suoraan perusmuotohakemistosta. Jos hakemistosta ei löydy hakusanan kanssa täsmäävää perusmuotoa, tehdään lisätarkistukset.

Näiden kahden tavan eroa havainnollistetaan tässä hakusanalla *puhelin*. Vaihtoehdossa a) se syötetään ensin perusmuoto-ohjelmalle, joka löytää kaksi tulkintaa: hakusana on perusmuoto puhelin-substantiivista tai taivutusmuoto puhella-verbistä. Koska puhelin-perusmuoto on sanakirjassa, sitä käytetään hakusanana. Jos yhtään perusmuototulkintaa ei olisi löytynyt, olisi tarkistamista jatkettu.

Vaihtoehdossa b) käyttäjän antama hakusana puhelin etsitään suoraan hakemistosta. Koska se löytyy hakemistosta, voidaan olettaa, että kaikki puhelin-sanan muodot on tallennusvaiheessa tunnistettu oikein ja kysely siten voi edetä normaalisti. Tässä ei siis tarvitse ottaa kantaa puhelin-sanan tyyppisiin monitulkintaisuuksiin, mikä tekee kyselyn suorittamisen siinä suhteessa nopeammaksi. Vain niissä tapauksissa, kun hakemistosta ei löydy yhtään hakusanan kanssa täsmäävää hakemistosanaa, hakusana tutkitaan tarkemmin. Epäonnistuminen voi johtua yksinkertaisesti siitä, että kyseinen sana ei todellakaan ole esiintynyt teksteissä. On kuitenkin varmistettava, ettei syynä ole se, että sana on tuntematon ja puuttuu perusmuoto-ohjelman sanakirjasta, tai että hakija olisi antanut hakusanan taivutusmuodossa.

Vaihtoehdon b) puutteena on, että jos hakemistosta löytyy hakusanan homografi, kyselyn tuloksena ei saada tyhjää joukkoa ja korjauskyselyt jäävät tekemättä. Jos esimerkiksi haetaan sanaliittoa Sri Lanka ja hakemistosta löytyy jälkimmäisen osan sananmuotohomografi lanka, tämä hakusana oletetaan tunnetuksi. Niinpä ei tehdä tarkistuskyselyä tunnistamattomien sananmuotojen hakemistosta eikä löydetä siellä olevia Lankassa, Lankasta jne. -muotoja. Tämän seurauksena hakutuloksen saanti alenee verrattuna perinteisellä tavalla tehtyyn kyselyyn, jossa käytetään hakijan katkaisemia hakusanoja. Tosin on huomattava, että vaihtoehto a), hakusanan tarkistaminen ennen varsinaista kyselyä, ei tässä tapauksessa toimi sen paremmin, sillä jos erisnimi Sri Lanka puuttuu sanakirjasta, sanaliiton loppuosa tässäkin tapauksessa tulkitaan lanka-sanaksi eikä lisätarkistuksia tehdä.

Hakujärjestelmän toteutuksesta riippuu, onko hakusanat välttämätöntä tarkistaa ennen kyselyn suorittamista vai tarkistetaanko ne vasta, mikäli hakemistosta ei löydy yhtään hakusanan kanssa täsmäävää hakemistosanaa. Jos esimerkiksi yhdyssanat on tallennusvaiheessa hajotettu osiinsa ja hakemis-

toon on tallennettu pelkästään nämä osat muttei yhdyssanaa kokonaisuudessaan, syöte pitää aina tutkia ennen kyselyn suoritusta, jotta yhdyssanat havaitaan ja puretaan osiinsa. Tämä osittaminen tulisi tehdä automaattisesti, samaa sanakirjaa ja säännöstöä käyttäen kuin tallennusvaiheessa, jotta hakusanat ovat varmasti samassa muodossa kuin hakemistosanat. Koska FULLTEXT-projektin perusmuotohakemistoihin oli tallennettu yhdyssanat myös kokonaisina, syötteen tarkistus toteutettiin vaihtoehdon b) mukaisesti eli vasta, mikäli kyselyn tuloksena ei saatu yhtään dokumenttia. - Myöskään osittaistäsmäyttävässä hakujärjestelmässä vaihtoehtoja ei ole, vaan hakusanat on aina tarkistettava, koska tulosjoukko ei siinä koskaan ole tyhjä.

7.6 Testikyselyjen laatiminen

7.6.1 Hakupyyntöjen keräys ja valinta

Tutkimuksessa käytetyt hakupyynnöt valikoitiin neljästä eri kysymyssarjasta. Näistä kaksi oli perinteiseen, manuaaliseen lehtileikearkistoon tehtyjä aitoja hakupyyntöjä, toiset kaksi sarjaa taas erityisesti tekstitiedonhaun kokeiluja varten tuotettuja, ei-aitoja hakupyyntöjä. Näissä neljässä kysymyssarjassa oli hakupyyntöjä yhteensä 509 kappaletta.

Ensimmäinen kysymyssarja oli Helsingin Sanomien leikearkiston arkistohakupyyntöjä, jotka Tarja Hjorth keräytti keväällä 1986. Näitä toimittajien aitoja hakupyyntöjä kerättiin yhteensä 99 kappaletta. Muistiinmerkittyjen kysymysten avulla tutkittiin, millaista tietoa Helsingin Sanomien leikearkistosta halutaan ja miten haluttu tieto löydetään. (Hjorth 1986)

Toinen kysymyssarja koostui hakupyynnöistä, joita Helsingin Sanomien toimittajat keksivät tekstitietokannan kokeilua varten. Näitä hakupyyntöjä saatiin kaikkiaan 173 kappaletta vuonna 1987. Arkistokokeilussa käytetty tietokanta sisälsi kotimaan, ulkomaan ja talousosastojen artikkeleita ajanjaksolta 26.5. - 29.6.1987, yhteensä 2 539 kappaletta. Hakupyynnöt olivat kyseisten osastojen toimittajien keksimiä. (Hjorth 1987a)

Kolmas sarja, 30 hakupyyntöä, saatiin Jaana Kristensenin pro gradu –tutkimuksesta. Aamulehden taloustoimittajia oli pyydetty keksimään kysymyksiä, joita he esittäisivät haettavaksi tekstitietokannasta. (Kristensen 1989, s. 37 - 38)

Neljäs hakupyyntösarja kerättiin FULLTEXT-projektia varten Helsingin Sanomien arkistosta syksyn 1990 aikana. Lomakkeen keräystä varten laati Tar-

ja Hjorth. Keräyksen tuloksena merkittiin muistiin 207 hakupyyntöä. Nämä hakupyynnöt olivat Helsingin Sanomien ja Ilta-Sanomien toimittajien tekemiä aitoja, perinteiseen leikearkistoon kohdistuvia hakupyyntöjä.

Näistä eri tahoilta kerätyistä hakupyynnöistä karsittiin aluksi pois sellaiset, jotka olivat sidoksissa johonkin erityiseen ajankohtaan tai paikkaan ja jotka siksi eivät olleet käyttökelpoisia testiaineistona. On epätodennäköistä (joskaan ei mahdotonta), että esimerkiksi syksyn ylioppilaskirjoituksista tai vuonna 1985 esitetystä Vanha koulukaveri -kuunnelmasta löytyy artikkeleita tietokannasta, joka kattaa vuoden 1990 ensimmäiset kolme kuukautta.

Nimet, erityisesti henkilönnimet ovat tyypillisiä lehtiarkiston hakupyyntöjä. Esimerkiksi Helsingin Sanomien syksyllä 1990 kerätyistä 207 kysymyksestä noin 80:n aiheena oli henkilönnimi eli niiden osuus kysymyksistä oli lähes 40 prosenttia. Jos hakujärjestelmässä voidaan käyttää läheisyysoperaattoreita, henkilönnimet eivät sinänsä ole erityisen hankalia haettavia. Ongelmia tuottavat tapaukset, joissa haettu henkilönnimi - tai yleensä erisnimi - on jokin muutenkin yleisesti esiintyvä sana (Sorsa, Meri, Sato) taikka homografinen jonkin yleisesti esiintyvän sanan tai sanan taivutusmuodon kanssa (Elin, Olin, Sulin).

Toinen nimiin liittyvä ongelmatyyppi ovat tietyssä taivutusmuodossa esiintyvät tai muuten poikkeavasti taipuvat nimet, kuten Yhtyneet Paperitehtaat tai Tuntematon sotilas. Mikäli dokumenteissa esiintyneet sanat on tallennusvaiheessa palautettu perusmuotoonsa, näitä nimiä ei löydetäkään suoraan hakemistosta vakiomuodossaan. Hakijan on kuitenkin luontevampaa syöttää tällaiset ilmaukset hakujärjestelmään tavanomaisessa esiintymismuodossaan kuin perusmuotoisina (yhtyä paperitehdas). Sitä paitsi hakija ei käytännössä voi tietää, miten kyseiset sanat on tallennusvaiheessa tulkittu ja missä muodossa tallennettu hakemistoon, koska se riippuu perusmuoto-ohjelman sanakirjasta ja säännöistä.

Edellämainittujen ongelmatyyppien eli homografisten tai taivutusmuodossa esiintyvien hakusanojen lisäksi hakupyynnöistä etsittiin sanoja, jotka olisivat johdos- tai yhdyssanalaajennosten kannalta kiinnostavia. Tämän valikoinnin perusteella tutkimusaineistona olleista neljän kysymyssarjan noin 500 hakupyynnöstä jäi jäljelle 143 mahdollista hakupyyntöä, joista tehtiin alustavat kyselyt.

Tutkimuksen tekijä poimi hakupyynnöistä sopivaksi katsomansa sanat ja tuotti niistä katkaistut hakusanat. Näistä muodostettiin kyselyt, joiden käyt-

tökelpoisuus testattiin suorittamalla kysely taivutusmuotohakemistosta. Hakutulosten perusteella karsittiin pois kyselyt, joihin ei saatu vastaukseksi yhtään dokumenttia tai vain muutama dokumentti, jotka eivät muodollisin perustein olleet käyttökelpoisia. Tällaisia olivat esimerkiksi kyselyt, joiden tulokseksi saatiin vain tapahtumakalentereita tai radio-ohjelmalistauksia. Periaatteessa radio-ohjelmia yms. aineistoa ei tallenneta Aamulehden elektroniseen arkistoon, mutta käytännössä tällaisiakin artikkeleita löytyi tutkimusaineistosta. Nämä artikkelit voitiin ilman sisällön analyysiäkin päätellä jo lehden osaston perusteella epärelevanteiksi.

Alustavan perinteisen kyselyn tuloksena tuli saada vähintään kaksi dokumenttia, jotta kysely kelpuutettiin jatkoon. Tällä pyrittiin siihen, että testikyselyjen tulokset eivät olisi tulosjoukkojen pienuuden vuoksi liian sattumanvaraisia.

Alustavien testikyselyjen tuottamien tulosten perusteella valittiin lopullisiin testauksiin 30 hakupyyntöä. Varsinaisten testikyselyjen (vakiokyselyt) lisäksi keksittiin itse vielä 15 hakupyyntöä, joissa pyrittiin selvittämään homografisen tai muuten kielellisesti hankalan ilmauksen analysoitumista. Nämä olivat lehtiarkistojen asiantuntijoiden suullisesti mainitsemia ongelmatapauksia sekä homografiaa käsittelevän kielitieteellisen kirjallisuuden avulla keksittyjä ongelmasanoja (Laalo 1990). Lopullisiin testauksiin valitut hakupyynnöt on lueteltu liitteessä 1.

7.6.2 Hakusanojen valinta

Testeihin valittujen 45 hakupyynnön pohjalta siis laadittiin 30 vakiokyselyä ja 15 ongelmakyselyä. Ensinmainituissa tutkittiin erityisesti kyselyjen laajentamista johdoksilla ja yhdyssanoilla, jälkimmäisessä puolestaan homografisten tai taivutusmuodossa esiintyvien hakusanojen käyttäytymistä.

Aluksi tutkija laati perinteiset kyselyt, joissa käytettiin katkaistuja hakusanoja. Tässä ajateltiin tilannetta, jossa hakija on ammattitaitoinen ja osaa käyttää hakujärjestelmän tarjoamia mahdollisuuksia hyväkseen, kuten katkaista hakusanan sopivimmasta kohdasta. Tavoitteena oli, että hakusanat ovat muodoltaan samanlaisia kuin silloin, kun dokumentteja etsitään todellisista, tuotantokäytössä olevista tekstitietokannoista.

Vakiokyselyt kuitenkin poikkesivat tositilanteessa käytettävistä kyselyistä siinä suhteessa, että niissä käytettiin pelkästään hakupyynnöissä esiintyneitä sanoja ja näiden johdos- ja yhdyssanavariaatioita. Vakiokyselyjä ei siis laajennettu hakusanojen synonyymeillä, rinnakkaisilmauksilla tai ylä- ja ala-

käsitteillä. Tiedonhaun ammattilaisethan pyrkivät löytämään hakusanoille tarpeen mukaan myös vaihtoehtoisia ilmauksia, koska haun aihe - erisnimiä lukuunottamatta - vain harvoin voidaan ilmaista tyhjentävästi yhdellä ilmauksella. Esimerkiksi *kauppa*-hakusanan rinnalla voidaan käyttää myös hakusanoja *myymälä*, *liike* jne. Jos kyselyihin (siis kaikkiin kyselyversioihin) olisi lisätty myös tämäntyyppiset vaihtoehtoiset ilmaukset, tulosjoukoista olisi ollut vaikeaa arvioida, miten nimenomaan morfologisten tulkintaohjelmien soveltaminen vaikuttaa hakutuloksiin.

Samaten testijärjestelyissä poikettiin normaalista tiedonhakutilanteesta siinä, että kyselyt suoritettiin vain kerran eikä niitä enää muokattu hakutulosten (relevanssipalautteen) perusteella. Normaalistihan tiedonhaku on vuorovaikutteinen prosessi, jossa hakija yrityksen ja erehdyksen kautta korjaa kyselyään periaatteessa niin kauan, että tulosjoukko sisältää vain toivotunlaisia dokumentteja (Swanson 1977). Käytännössä kuitenkin käytettävissä olevat resurssit yms. ratkaisevat, kuinka pitkään hakija voi jatkaa kyselyn muokkausta. Aitojen hakijoiden puuttuminen on ollut tyypillistä tiedonhakututkimukselle: massiivisessa TREC-hankkeessakin on todettu, että perinteisen järjestelmäsuuntautuneen tutkimuksen ehdoilla rakennetussa tutkimusjärjestelmässä on vaikea tutkia sitä, miten hakijat aidosti muokkaavat kyselyjä hakutulosten perusteella. TREC-hankkeessa tämän vuorovaikutuksen tutkiminen on vielä alkuvaiheessaan, vaikka tutkimusaluetta varten on perustettu oma osionsakin (interactive track) vuonna 1995. (Beaulieu et al. 1996; Sparck Jones 2000)

Kaikkia hakupyynnössä esiintyneitä sanoja ei käytetty tämän tutkimuksen testikyselyissä. Aina ei myöskään käytetty hakupyynnön sanoja sellaisinaan, vaan hakusanojen valinta ja katkaisu pyrittiin tekemään mahdollisimman järkevästi, ns. hyvää ammattikäytäntöä noudattaen. Esimerkiksi lomaosakebisnes-sanaa ei kannattanut käyttää sellaisenaan, koska bisnes olisi rajoittanut kyselyä niin, ettei tulokseksi olisi saatu yhtään artikkelia. Tällaisissa tapauksissa hakusanasta käytettiin muotoa, jota todellisessakin hakutilanteessa käytettäisiin (*lomaosak**).

Yhdyssana hajotettiin osiinsa, mikäli osista voitiin muodostaa mielekäs sanaliitto: esimerkiksi sanaliitto autojen verotus on käytännössä synonyyminen autoverotus-sanan kanssa. Osiin jakamisen ohjenuorana oli se, miten

tiedonhaun ammattilainen jakaisi yhdyssanat - esimerkiksi seurakunta-sanaa informaatikot (tai satunnaisetkaan hakijat) tuskin hajottaisivat. [2]

Mikäli kyselyssä oli useampia hakusanoja, kyselyistä tuotettiin kaksi versiota: toisessa hakusanat kytkettiin toisiinsa JA-operaattorilla ja toisessa läheisyysoperaattorilla (käytännössä virkeoperaattorilla).

Perinteisten kyselyjen vastaavuus asiantuntevan ammattikäytännön kanssa varmistettiin antamalla ne alan asiantuntijoiden eli kolmen kokeneen VTT:n informaatikon tarkastettaviksi, minkä jälkeen kyselyt korjattiin informaatikoilta saadun palautteen perusteella. Kaikki kolme informaatikkoa olivat informaatiopalvelukurssin käyneitä ja kaikkien kolmen ensisijaisena työtehtävänä oli tiedonhakujen tekeminen. Yksi koehenkilöistä oli ollut informaatikon tehtävissä suunnilleen yhden vuoden ja kaksi muuta noin seitsemän vuoden ajan.

Informaatikot esittivät vain muutamia korjausehdotuksia. Korjaukset olivat joko ehdotus jonkin hakusanan katkaisukohdan muuttamiseksi tai jonkin hakusanan lisäämiseksi kyselyyn. Kaikki korjausehdotukset olivat sellaisia, että niitä ehdotti vain yksi kolmesta informaatikosta. Korjaukset tehtiin, mikäli syynä oli tutkijan virhe tai hakusanan unohtaminen, informaatikon väärinkäsityksestä johtuvat korjausehdotukset jätettiin tekemättä.

7.6.3 Kyselytyyppien muodostaminen eri tutkimusympäristöjen vertailua varten

Kun perinteiset kyselyt olivat valmiit ja korjattu informaatikoilta saadun palautteen mukaisiksi, tutkimuksen tekijä laati niiden pohjalta vastaavat perusmuotoisia hakusanoja sisältävät kyselyt. Näiden suhteen oletettiin, että hakija tuntee hakujärjestelmän toimintaperiaatteet ja tietää muun muassa, mikä on hakusanan perusmuoto tai onko hakusana perussana vai yhdyssana. Koska VTT:n informaatikoilla ei ollut perusmuotoisista hakusanoista enempää käytännön kokemusta kuin tutkijalla itsellään, näitä kyselyjä ei enää annettu asiantuntijaraadin tarkastettaviksi.

Tutkimuksen lähtökohtana oli verrata kyselyjä, joissa hakija on antanut hakusanat perusmuodossa, perinteisiin kyselyihin, joissa hakija on katkaissut hakusanat itse. Mutta pelkästään tämä ei riitä. Jos vertaillaan suora-

[2] Tätä voidaan verrata perinteisissä kaupallisissa hakujärjestelmissä tarjottuun mahdollisuuteen katkaista hakusana: hakija voi katkaista hakusanan, mutta hän voi myös jättää sen katkaisematta; yhdyssanojenkin suhteen hakijalle tulisi antaa mahdollisuus jättää yhdyssana jakamatta, vaikka oletusarvona olisikin, että yhdyssanat jaetaan osiinsa.

viivaisesti vain näitä kahta eri vaihtoehtoa, eri tutkimusympäristöjen kyselyt eivät ole yhteismitallisia: Kun taivutusmuotohakemistosta haetaan katkaistuilla hakusanoilla (merkkijonokaavioilla), hakusana poimii kaikki tietyllä merkkijonolla alkavat hakemiston merkkijonot, olivatpa ne sitten perussanoja tai yhdyssanoja. Merkkijonokaavio *vero** palauttaa muun muassa hakemiston merkkijonot vero, veron, verolaki, ja veronkierto. Jos katkaistu hakusana on lyhyt, se saattaa palauttaa myös hakusanan johdoksia (verotus, verottaja, verottaminen jne.). Sen sijaan perusmuotohakemistosta haettaessa hakusana *vero* täsmää vain hakemiston merkkijonoon vero. Johdokset ja yhdyssanat on haettava erikseen, jos nekin halutaan löytää.

Kyselyjä oli siis muokattava siten, että hakusanojen erilainen käyttäytyminen eri tutkimusympäristöissä tulisi otetuksi huomioon. Jotta tämä olisi saatu mahdollisimman selkeästi ja systemaattisesti esiin, kyselyistä tuotettiin joukko erityyppisiä muunnelmia, **kyselytyyppejä**[3]. Eri kyselytyypeissä yhdyssanat ja johdokset esiintyivät eri tavoin (taulukko 1).

Suppein kyselytyyppi on se, joka sisältää vain hakupyynnössä esiintyneiden sanojen perusmuodot:

HAE *autoverotus*	(hakupyyntö numero 13)
HAE *lapsi* JA *mainonta*	(hakupyyntö 9)

Tämän suppeimman kyselytyypin, **peruskyselyn**, symbolina on **A**.

Mikäli hakusanalle oli keksittävissä johdoksia tai se itse oli johdos, jolle löytyi läheinen kantasana, peruskyselyä laajennettiin lisäämällä nämä johdokset, kantasana tai molemmat TAI-operaattorilla kytkettyinä mukaan kyselyyn:

HAE *autoverotus* TAI *autovero*

Kantasanan ja johdosten muodostaman sanaryhmän yhteisnimitys on **johdosperhe** ja hakusanan ja sen muun johdosperheen perusmuodot sisältävää kyselyä kutsutaan **johdoskyselyksi**. Sen symbolina on jatkossa **AB**.

Kunkin kyselyn mahdolliset johdos- ja kantasanalaajennukset suunnitteli tutkija itse ennen kuin kyselyjä käytännössä suoritettiin. Tällä haluttiin välttää se, että hakemistossa olevat hakemiston merkkijonot vaikuttaisivat kyselyjen suunnitteluun ja siten mahdollisesti vinouttaisivat tutkimusta. Johdos-

[3] Seuraavassa esitetty jaottelu on sama kuin VTT:n julkaisussa (Alkula & Honkela 1992), vaikka kyselytyyppien tunnukset on merkitty eri tavalla - vain notaatio on muuttunut.

ten ideoinnissa käytettiin hyväksi muun muassa suomen kielen kieliopin johdinluetteloita (Karlsson 1987). - Näin ollen oli täysin mahdollista, että hakemistosta ei löytynyt yhtään johdosperheen jonkin tai joidenkin jäsenten kanssa täsmääviä hakemistosanoja. Tällainen käytäntö vastaa tilannetta, jossa kyselyä automaattisesti laajennetaan ennaltalaaditun hakutesauruksen avulla - edellyttäen, että johdosperhe olisi tällaisessa hakutesauruksessa erikseen valittavissa (vrt. Kristensen 1989; 1990; 1992; 1993).

Kyselytyypit A ja AB olivat käytössä vain tutkimusympäristöissä T4 (perusmuotohakemisto), T5 (ositettu perusmuotohakemisto) ja T6 (kaksoishakemisto). Taivutusmuotohakemistosta ei kannata hakea pelkkiä perusmuotoja, koska silloin tulosjoukosta puuttuisivat kaikki dokumentit, joissa hakusana on esiintynyt pelkästään taivutusmuodossa. Esimerkiksi genetiivimuoto voi esiintyä tekstissä huomattavasti useammin kuin nominatiivimuoto (katso luku 8.6).

Seuraavassa kyselytyypissä laajennettiin haun alaa dokumentteihin, joissa oli esiintynyt hakusanan perusmuoto tai sellainen yhdyssana, jonka osana hakusana on. Tutkimusympäristöissä T2 (automaattinen katkaisu), T3 (seulonta) ja T4 (perusmuotohakemisto) tällainen **yhdyssanakysely** (symboli **AC**) muodostettiin syöttämällä ensin perusmuodossa olevat hakusanat taivutusvartaloita tuottavalle ohjelmalle. Vartalo-ohjelman tuottamien vartaloiden perään liitettiin automaattisesti hakujärjestelmän katkaisumerkki ja näin saadut hakusanat sijoitettiin kyselyyn alkuperäisen hakusanan tilalle. Näin saatiin dokumentit, joissa esiintyi hakusanalla alkava yhdyssanat. Siten kysely oli esimerkiksi seuraavanlainen:

> HAE *autoverotus** TAI *autoverotuks** TAI *autoverotude** TAI *autoverotut**

Esimerkin kaksi jälkimmäistä vartaloa ovat seurausta vartalo-ohjelmien yligeneroinnista. Koska kaikkien sanojen taivutusluokkaa ei voi yksiselitteisesti päätellä sanahahmosta, tuotetaan vartalot laajimman mahdollisen taivutusluokan mukaan. Esimerkiksi rakkaus-sanan taivutusluokkaan kuuluvat edellä esitetyn esimerkin mukaiset neljä vartaloa, sen sijaan verotus-sanan kannalta kaksi vartaloista on ylimääräisiä. Kyselyä suoritettaessa tällaisista ylimääräisistä taivutusvartaloista ei juuri ollut haittaa; niillä haettaessa vain saatiin tulokseksi tyhjä joukko.

Taulukko 1. Eri kyselytyypit ja esimerkki kussakin kyselytyypissä käytetyistä hakusanoista. Täydelliset kyselyt on esitetty liitteessä 2.

A symboloi suppeaa perusmuodot sisältävää kyselyä, B kyselyn laajennusta johdosperheellä ja C laajennusta yhdyssanoilla. Pienet kirjaimet a, b ja c viittaavat yhdyssanojen osittamiseen ja kyselyn laajentamiseen vastaavasti kuin edellä (a laajennusta osien perusmuodoilla, b osien johdosperheellä ja c yhdyssanan kaikilla osilla).

KYSELYTYYPIT	**HAKUSANAT** (esimerkkinä autoverotus)
Perinteinen yhdistelmäkysely (ABC)	*autovero**
Perinteinen osien yhdistelmäkysely (ABCabc)	*auto** JA *vero**
Peruskysely (A)	*autoverotus*
Johdoskysely (AB)	*autoverotus* TAI *autovero* TAI *autoverottaminen*
Yhdyssanakysely (AC)	*autoverotus* TAI *autoverotus-* TAI *-autoverotus-* TAI *-autoverotus*
Yhdistelmäkysely (ABC)	*autoverotus* TAI *autoverotus-* TAI *-autoverotus-* TAI-*autoverotus* TAI *autovero* TAI *autovero-* TAI *-autovero-* TAI *-autovero* TAI *autoverottaminen* TAI *autoverottaminen-* TAI *-autoverottaminen-* TAI *-autoverottaminen*
Osien peruskysely (Aa)	*autoverotus* TAI (*auto* JA *verotus*)
Osien johdoskysely (ABab)	*autoverotus* TAI *autovero* TAI *autoverottaminen* TAI *auto* JA (*verotus* TAI *vero* TAI *verottaja* TAI *verottaminen* TAI *verottaa*)
Osien yhdyssanakysely (ACac)	(*auto* TAI *auto-* TAI *-auto-* TAI *-auto*) JA (*verotus* TAI *verotus-* TAI *-verotus-* TAI *-verotus*)
Osien yhdistelmäkysely (ABCabc)	(*auto* TAI *auto-* TAI *-auto-* TAI *-auto*) JA (*verotus* TAI *verotus-* TAI *-verotus-* TAI *-verotus* TAI *vero* TAI *vero-* TAI *-vero-* TAI *-vero* TAI *verottaja* TAI *verottaja-* TAI *-verottaja-* TAI *-verottaja* TAI *verottaminen* TAI *verottaminen-* TAI *-verottaminen-* TAI *-verottaminen* TAI *verottaa* TAI *verottaa-* TAI *-verottaa-* TAI *-verottaa*)

Tutkimusympäristössä T5 eli ositetussa perusmuotohakemistossa yhdyssanakysely toteutettiin siten, että hakija lisäsi itse manuaalisesti hakusanan perusmuotoon yhdyssanojen eri osia symboloivat katkaisumerkit. Tämä olisi periaatteessa helppoa toteuttaa automaattisestikin siten, että hakija merkitsee perusmuotoiseen hakusanaan määrätyn koodin, jonka perusteella hakujärjestelmä osaa lisätä kyselyyn perusmuodon rinnalle myös yhdyssanan eri osat asianmukaisin katkaisumerkein varustettuina. T5-ympäristössä AC-kyselyt olivat siten seuraavanlaisia:

> HAE *autoverotus* TAI *autoverotus-* TAI *-autoverotus-* TAI *-autoverotus*

Kysely, jossa käytetään kokonaisia yhdyssanoja hakusanoina sellaisenaan, edellyttää tietenkin, että hakemistoon on tallennettu kokonaiset yhdyssanat. Jos hakemistoon on tallennettu vain osat, joista yhdyssanat muodostuvat (tässä tapauksessa siis *auto-*, *-verotus*; vrt. luku 7.2.), hakuvaiheen alussa pitäisi ensin suorittaa ohjelma, joka automaattisesti pilkkoo hakijan antamat hakusanat osiinsa samalla tavalla kuin hakemistosanat on pilkottu tallennusvaiheessa.

Ositetun perusmuotohakemiston kautta voidaan siis helposti hakea dokumentteja, joissa hakusana on esiintynyt yhdyssanojen alku-, keski- ja loppuosina. Taivutusmuotohakemistosta ja pelkät (osittamattomat) perusmuodot sisältävästä hakemistostahan voidaan helposti poimia vain yhdyssanat, joissa hakusana on alkuosana.

Yhdistelmäkyselyssä haetaan dokumentteja, jotka sisältävät hakusanan tai sen johdosperheen jäsenen perusmuodon tai sellaisen yhdyssanan, jonka osana alkuperäinen hakusana tai jokin sen johdosperheen jäsen on. Yhdistelmäkyselyn tulosjoukko ei siis ole vain johdoskyselyn (AB) ja yhdyssanakyselyn (AC) tuottamien tulosjoukkojen unioni, vaan siinä on niihin verrattuna aidosti uusia hakusanoja. Yhdistelmäkyselyn symbolina on jatkossa **ABC**.

Tutkimusympäristössä T1 käytettiin vain yhdistelmäkyselyä ABC, ei lainkaan sitä suppeampia kyselytyyppejä. Perinteisessä yhdistelmäkyselyssä hakija katkaisi hakusanat siten, että katkaistuilla hakusanoilla voitiin mahdollisimman hyvin kattaa hakusanan taivutusmuotojen ja yhdyssanojen lisäksi myös sen johdosperheen jäsenet ja näitä sisältävät yhdyssanat.

Tutkimusympäristössä T5 perusmuotoisiin johdosperheen jäseniin lisättiin yhdyssanojen osia symboloivat katkaisumerkit samaan tapaan kuin edellä

yhdyssanakyselyn (AC) yhteydessä selostettiin. Näin saadut hakusanat lisättiin kyselyyn TAI-operaattorilla kytkettyinä:

HAE *autoverotus* TAI *autoverotus-* TAI *-autoverotus-* TAI *-autoverotus*
TAI *autovero* TAI *autovero-* TAI *-autovero-* TAI *-autovero*

Muissa tutkimusympäristöissä (T2, T3, T4) yhdistelmäkysely toteutettiin niin, että vartalo-ohjelmalle syötettiin varsinaisen hakusanan lisäksi myös sen johdosperheen muut jäsenet. Vartalo-ohjelman tuottamat uudet taivutusvartalot lisättiin kyselyyn alkuperäisen hakusanan rinnalle TAI-operaattorilla yhdistettynä:

HAE *autoverotus** TAI *autoverotuks** TAI *autoverotude** TAI *autoverotut** TAI *autovero**

Tässä esimerkkitapauksessa huomataan, että yksi johdosperheen jäsenistä eli hakusana *autovero** kattaa myös kaikki muut vaihtoehdot. Tällaisissa tapauksissa kysely käytännössä toteutettiinkin niin, että haettiin vain tällä yhdellä johdosperheen jäsenellä.

Seuraava peruskyselystä laajennettu vaihtoehtoinen kyselytyyppi oli **osien peruskysely**. Mikäli hakusana oli yhdyssana, jonka osista voitiin tutkijan mielestä ja informaatikkoraadin vahvistamana muodostaa mielekäs sanaliitto, peruskyselyä laajennettiin perusmuotoisilla yhdyssanan osilla. Esimerkiksi autoverotus ei välttämättä esiinny tekstissä juuri yhdyssanana, vaan ositetussa muodossa autojen verotus. Osien peruskyselyn symbolina on jatkossa **Aa**[4].

HAE *autoverotus* TAI (*auto* JA *verotus*).

Osien johdoskyselyssä puolestaan yhdyssanan ja sen johdosperheen (eli johdoskyselyn hakusanojen) rinnalle kyselyyn lisättiin hakusanoiksi myös yhdyssanan osat ja yhdyssanan osien johdosperheiden jäsenet perusmuodossa. Osien johdoskyselyn symbolina on **ABab**.

HAE *autoverotus* TAI *autovero* TAI (*auto* JA (*verotus* TAI *vero*))

Osien peruskyselyä ja osien johdoskyselyä käytettiin vain T4- ja T5-tutkimusympäristöissä, joissa hakemiston sanat oli perusmuotoistettu. Kuten

[4] Notaatiossa isojen ja pienten kirjainten erona on, että isot kirjaimet symboloivat hakusanoja (kyselyjä) ja niiden variaatioita kokonaisina (yhdys)sanoina, mutta pienet kirjaimet symboloivat yhdyssanojen jakamista osiin. Näin esimerkiksi osien peruskyselyn pieni a viittaa siihen, että peruskyselyn A sisältämät yhdyssanat on "pienitty" perussanoiksi eli jaettu osiinsa.

edellä on perusteltu, taivutusmuotohakemistosta ei kannata hakea pelkillä perusmuotoisilla (katkaisemattomilla) hakusanoilla.

Osien yhdyssanakyselyllä haettiin dokumentteja, joissa esiintyisi yhdyssanoja, joiden jonain osana alkuperäisen hakusanan osat ovat. Esimerkiksi autoverotus ei välttämättä esiinny nimenomaan tällaisena yhdyssanana, vaan sen osat voivat olla eri sanoissa, kuten kuorma-*autojen* *verotus*käytäntö. Osien yhdyssanakyselyn symbolina on **ACac**.

Tutkimusympäristöissä T2 (automaattinen katkaisu), T3 (seulonta) ja T4 (perusmuotohakemisto) osien yhdyssanakysely toteutettiin syöttämällä yhdyssanan osien perusmuodot taivutusvartaloita tuottavalle ohjelmalle. Näin saatujen taivutusvartaloiden perään liitettiin automaattisesti hakujärjestelmän katkaisumerkki ja ne lisättiin alkuperäisen hakusanan rinnalle kyselyyn TAI-operaattorilla kytkettyinä, esimerkiksi:

> HAE *autoverotus** TAI *autoverotuks** TAI *autoverotude** TAI *autoverotut** TAI (*auto** JA *verotus**).

Sen sijaan tutkimusympäristössä T5 (ositettu perusmuotohakemisto) kysely toteutettiin niin, että hakija lisäsi yhdyssanan osiin manuaalisesti katkaisumerkit, jotka symboloivat yhdyssanojen eri osia:

> HAE (*auto* TAI *auto-* TAI *-auto-* TAI *-auto*)
> JA (*verotus* TAI *verotus-* TAI *-verotus-* TAI *-verotus*)

Toisin kuin edellä mainituissa kolmessa tutkimusympäristöissä, T5-tutkimusympäristön ositetusta perusmuotohakemistosta ei tarvinnut erikseen hakea kokonaista yhdyssanaa autoverotus eri variaatioineen. Sehän löytyi osia hakusanoina käytettäessä (*auto-* JA *-verotus -> autoverotus*).

Teknisesti olisi ollut mahdollista laatia kyselyjä, joissa hakusanoina käytetään pelkästään tiettyjä yhdyssanan osia, kuten pelkästään yhdyssanan lopussa esiintyviä verotus-sanoja. Testikyselyistä ei kuitenkaan tuotettu tällaisia lisämuunnelmia, koska niiden laatiminen, suorittaminen ja analysointi olisi vaatinut enemmän resursseja kuin FULLTEXT-projektissa oli mahdollista käyttää.

Osien yhdistelmäkysely (**ABCabc**) on laajin mahdollinen kyselytyyppi. Siinä osien yhdyssanakyselyä laajennetaan yhdyssanan osien johdosperheen jäsenillä vastaavalla tavalla kuin osien peruskyselystä laajennettiin osien johdoskysely. Osien yhdistelmäkyselyllä haetaan ne dokumentit, joissa esiintyy alkuperäinen hakusana, sen johdosperheen jäsen taikka yhdyssana tai sanaliitto, joka sisältää hakusanan tai sen johdosperheen jäsenen osat.

Taulukko 2. Yhteenveto eri tutkimusympäristöissä käytetyistä kyselytyypeistä (tarkempi kuvaus esimerkkeineen liitteessä 2).

TUTKIMUS-YMPÄRISTÖ	**KYSELYTYYPIT**
T1	ABC ABCabc
T2	AC, ABC ACac, ABCabc
T3	AC, ABC ACac, ABCabc
T4	A, AB, AC, ABC Aa, ABab, ACac, ABCabc
T5	A, AB, AC, ABC Aa, ABab, ACac, ABCabc

Ympäristöissä T2 (automaattinen katkaisu), T3 (seulonta) ja T4 (perusmuotohakemisto) osien yhdistelmäkysely toteutettiin niin, että vartalo-ohjelmalle syötettiin hakusanan ja sen osien lisäksi myös hakusanan johdosperheen muut jäsenet ja johdosperheen osat.

> HAE *Wärtsilä** JA ((*tilintarkastus** TAI *tilintarkastuks** TAI *tilintarkastude** TAI *tilintarkastut** TAI *tilintarkastaminen** TAI *tilintarkastamis** TAI *tilintarkastaj**)
> TAI ((*tili** TAI *tile**) JA (*tarkastus** TAI *tarkastuks** TAI *tarkastude** TAI *tarkastut** TAI *tarkast** TAI *tarkastaminen** TAI *tarkastamis** TAI *tarkastaj**))).

Edellä esitetyssä esimerkissä yksi hakusanoista, *tarkast**, itse asiassa kattaa kaikki muut sen kanssa TAI-operaattorilla kytketyt hakusanat. Tällaisissa tapauksissa kyselyssä käytettiin vain tätä yhtä, eri vaihtoehdot kattavaa hakusanaa.

Tutkimusympäristössä T5 (ositettu perusmuotohakemisto) osien yhdistelmäkysely toteutettiin lisäämällä yhdyssanan eri osiin ja eri osien johdosperheen jäseniin katkaisumerkit, jotka symboloivat yhdyssanojen eri osia:

> HAE (*auto* TAI *auto-* TAI *-auto-* TAI *-auto*) JA (*verotus* TAI *verotus-* TAI *-verotus-* TAI *-verotus* TAI *vero* TAI *vero-* TAI *-vero-* TAI *-vero* TAI *verottaja* TAI *verottaja-* TAI *-verottaja-* TAI *-verottaja* TAI *verottaminen* TAI *verottaminen-* TAI *-verottaminen-* TAI *-verottaminen* TAI *verottaa* TAI *verottaa-* TAI *-verottaa-* TAI *-verottaa*).

7.7 Hakutulosten relevanssiarviot

Kun FULLTEXT-projektin kyselyt oli suoritettu, jokaisesta tulosjoukosta poistettiin aluksi sellaisten osastojen artikkelit, jotka olivat pelkästään erilaisia tapahtumakalentereita tai radio- ja televisio-ohjelmia. Karsinta tapahtui suoraan osaston perusteella, ilman varsinaista sisältöanalyysiä.

Jokaista 45 hakupyyntöä varten laadittiin taulukko, jonka riveille oli listattu kaikki eri kyselyillä siihen vastaukseksi saadut artikkelit (käytännössä artikkelin numero oli rivin otsikkona). Sarakeotsikoissa puolestaan listattiin tutkimusympäristöt ja näiden kyselytyypit. Taulukon sarakkeeseen tuli artikkelin kohdalle merkintä, mikäli artikkeli sisältyi sarakkeessa mainitun kyselytyypin tulosjoukkoon (esimerkkitaulukko liitteessä 6). Taulukon avulla voitiin verrata, miten eri kyselytyypit tietyn tutkimusympäristön sisällä keskenään erosivat ja toisaalta, miten eri tutkimusympäristöistä saadut tulosjoukot erosivat toisistaan.

Tutkimusympäristöjen ja kyselytyyppien välisiä eroja analysoitiin sekä määrällisesti että laadullisesti. Määrällisiä eroja tutkittiin käymällä läpi taulukot ja tutkimalla artikkelit, jotka eri tutkimusympäristöissä ja eri kyselytyypeillä käyttäytyivät eri tavoin (miksi jokin tietty artikkeli sisältyi tai ei sisältynyt tulosjoukkoon). Tutkija luki artikkelit läpi ja tarkasti, minkätyyppisiin artikkelissa esiintyneisiin sananmuotoihin eri hakuavaimet täsmäsivät (luku 9).

Tulosjoukkojen laadullinen arviointi perustui relevanssiarvioihin. Tässä tutkimuksessa relevanssitulkintojen mielivaltaisuus pyrittiin estämään määrittelemällä ennakolta tiedontarvitsijan rooli (toimittaja) ja tilanne (jutun kirjoittaminen tietystä aihepiiristä). Rajaus oli mielekäs, koska hakupyynnöt oli alunperin kerätty tällaisista tilanteista ja myös tutkimuksessa käytetyt testidokumentit oli poimittu tähän tarkoitukseen rakennetusta sanomalehtiarkistosta.

Relevanssiarviot tehneet kolme koehenkilöä (mies ja kaksi naista) olivat kaksi kertaa viikossa ilmestyvän lehden uutistoimittajia, joilla kaikilla oli usean vuoden työkokemus. Kyseessä oli niin sanottu asiantuntijaraati, eli koehenkilöt eivät olleet alkuperäisten hakupyyntöjen esittäjiä. Koeasetelma eli tausta-aineiston hankkiminen jutun kirjoittamista varten oli heille sinänsä tuttu tilanne, vaikka heidän omassa lehdessään ei tutkimuksen ajankohtana ollutkaan käytettävissä samanlaista elektronista arkistoa kuin Aamulehdessä. Vaikka relevanssin arvioijien tulisi ihanteellisessa tutkimusasetelmassa olla samoja henkilöitä kuin alkuperäisten tiedontarvitsijoiden, ei tällaiseen

ollut FULLTEXT-projektin tutkimusasetelmassa mahdollisuuksia (hakupyynnöt oli kerätty eri lähteistä ja ne oli annettu eri aikoina). Toisaalta on myös todettu, että asiantuntijoiden ja todellisten tiedontarvitsijoiden relevanssiarviot eivät juuri eroa toisistaan (Borlund & Ingwersen 1997; Borlund 2000), joten raadin käyttö relevanssin arviointiin oli tiedonhakututkimuksen käytäntöjen mukaista.

Arvioinnin valmisteluvaiheessa otettiin ensin suurista tulosjoukoista otanta. Suureksi tulosjoukoksi tutkimuksessa oli määritelty yli 75 dokumenttia sisältävä tulosjoukko (otantaperiaatteet selostetaan tarkemmin luvussa 7.8). Tämän jälkeen kunkin hakupyynnön kaikista eri kyselytyypeillä eri tutkimusympäristöistä saaduista tulosjoukoista otettiin yhdiste.

Artikkeliniput annettiin toimittajien arvioitavaksi evästettynä ohjeella, että heidän piti arvioida kunkin artikkelin käyttökelpoisuus tilanteessa, jossa heidän olisi kirjoitettava artikkeli hakupyynnön aihepiiristä (liite 3). Itse artikkeleihin ei merkitty, minkälais(t)en kysely(je)n tuloksena ne oli saatu, vaan toimittajat saivat nähtäväkseen vain kuvitteellisen arkistohaun lopputulokset. Arvioitavaksi annettiin ns. vakiokyselyjen tulokset. (Esimerkki artikkelista liitteessä 4.) Vakiokyselyitä oli kaikkiaan 30 ja niiden tuloksena saatuja artikkeleita annettiin arvioitavaksi yhteensä 1 488 kappaletta. Artikkeliniput annettiin kullekin koehenkilölle eri järjestyksessä, jotta koehenkilöiden rutinoitumiseen ja oppimiseen liittyviä vinoumia ei syntyisi, tai että ne ainakin jakautuisivat tasaisemmin koko aineiston osalle.

Alkuperäiset hakupyynnöt olivat usein lyhyitä, parin sanan mittaisia lauseita, jotka olisivat sallineet monenlaiset tulkinnat. Jotta tulkinnanvaraisuutta saataisiin vähenemään ja siten relevanssiarvioille yhtenäinen perusta, tutkija sepitti kunkin hakupyynnön tueksi taustatarinan, jossa hakupyynnön aihepiiriä ja tiedontarvetta selostettiin muutamalla virkkeellä (liite 1). Taustatarinan tarkoituksena oli auttaa toimittajaa hahmottamaan, minkätyyppiseen tiedontarpeeseen artikkeleita tarvitaan (Borlund & Ingwersen 1997).

Taustatarinaa ei tietenkään tarvita silloin, kun tiedontarvitsijat itse arvioivat dokumenttien relevanssin. Kun FULLTEXT-projekti aloitettiin, taustatarinan käyttö ei ollut yleistä eikä mallia voinut ottaa aiemmista tutkimuksista. Sittemmin hakupyyntöä täydentäviä kuvauksia (narratiivi) on otettu käyttöön muun muassa TREC-hankkeessa (Harman 1993). Myös Suomessa Sormusen (1994) tutkimuksessa oli tiedontarve kuvattu laajemmin kuin vain lyhyenä hakulauseena.

Kutakin vakiokyselyä varten oli lomake, johon oli merkitty tulosjoukkoon sisältyvien artikkelien tunnusnumerot. Toimittajat lukivat jokaisen artikkelin ja merkitsivät arviointilomakkeeseen artikkelin tunnuksen kohdalle, oliko artikkeli heidän mielestään hyvin, jonkin verran vai ei ollenkaan hyödyllinen, mikäli heidän tulisi kirjoittaa artikkeli hakupyynnön aihepiiristä. Klassisissa hakututkimuksissa relevanssiarvio tavallisesti oli kaksijakoinen, eli dokumentti joko oli relevantti tai sitten ei. Kolmitasoisena tämän tutkimuksen jaottelu siis otti myös osittaisrelevanssin huomioon (vrt. Spink et al. 1998; Borlund & Ingwersen 1997; Borlund 2000).

Artikkeli sai enemmistön mielipiteen mukaisen luokituksen. Eli jos kaikki kolme tai kaksi kolmesta toimittajasta oli antanut saman luokituksen, artikkelille annettiin tämä luokka. Asetelma siis oli samanlainen kuin Tenopirin ja Ron (1990, s. 111) Harvard Business Review -tutkimuksessa (HBR/O). HBR/O-tutkimuksessa relevanssiarvioinnin teki kolmen asiantuntijan raati ja enemmistön mielipide ratkaisi, minkä luokituksen yksittäinen artikkeli sai. FULLTEXT-projektissa arvioitujen artikkelien lukumäärä ja luokitus hakupyynnöittäin on esitetty liitteessä 5.

FULLTEXT-tutkimuksessa löytyi myös sellaisia artikkeleita, joista kaikki kolme arvioijaa olivat eri mieltä: yhdelle artikkeli oli erittäin käyttökelpoinen, toiselle jonkin verran ja kolmannelle ei ollenkaan. Tällöin artikkeli luokitettiin keskimmäiseen ryhmään eli näkemysten keskiarvon mukaisesti "jonkin verran käyttökelpoiseksi".

Toimittajille annetuissa ohjeissa ei puhuttu lainkaan artikkeleiden aiheenmukaisuudesta, vaan vain artikkeleiden hyödyllisyydestä tietyssä **tilanteessa** (liite 3). Tällainen määrittely sallii periaatteessa myös tapaukset, joissa artikkeli katsotaan hyödylliseksi, vaikka se ei käsittelisikään hakupyynnön aihetta. Oikeampi sanoitus kenties olisi ollut, onko artikkeli hyödyllinen tietyn **aiheen** kannalta. FULLTEXT-projektin koehenkilöt eivät kuitenkaan havainneet koeasetelmassa tätä relevanssiteoreettista ongelmaa - ehkä siksi, että tausta-artikkeleiden hankinta oli heille ammattikokemuksen kautta tuttu tilanne.

Toinen kysymyksenasettelun puute oli se, ettei toimittajille määritelty, millainen artikkeli heidän kuvitellussa tilanteessa pitäisi kirjoittaa: artikkelin pituutta, osastoa, tyyppiä tms. ei ollut määritelty. Artikkelin pituus tai sen kiireellisyys vaikuttaa siihen, kuinka paljon toimittaja ehtii tai haluaa etsiä tausta-aineistoa kirjoittamisen tueksi. Yksi koehenkilöistä tosin totesi arvioinnin jälkeen, ettei tämän määrittelyn puuttuminen ollut ongelma. Hänen

mielestään toimittajat käytännössä mieltävät "jutun" 1 - 3 konekirjoitusliuskan uutiseksi, mikäli sitä nimenomaisesti ei määritellä joksikin muuksi.

Kuten edellä jo todettiin, koejärjestelyjen puutteena oli myös se, että toimittajat eivät itse olleet alkuperäisiä tiedontarvitsijoita. Lyhyt hakupyyntö voi sallia liian monet tulkinnat, mutta toisaalta tutkijan kirjoittama taustatarina voi olla liian rajoittava ja tulkintoja liikaa ohjaava. Periaatteessa vaarana on myös se, että jos tutkija ennakolta tuntee testitietokannan sisällön, hän laatii tarinat sellaisiksi, että relevantteja dokumentteja löytyy varmasti. Suurissa tietokannoissa tämä lienee enemmän teoreettinen ongelma, koska kymmeniä tuhansia tai vieläkin useampia artikkeleita sisältävän tietokannan sisältöä ei voi yksityiskohtaisesti tuntea. Taustatarinan avulla toimittajien arviot kuitenkin ovat yhdenmukaisempia ja systemaattisempia kuin ilman minkäänlaista ohjausta. (Borlund & Ingwersen 1997).

Neljäs ongelma, joka tuloksia analysoitaessa paljastui, oli se, että kehyskertomuksesta huolimatta hakupyynnön aihepiiri saattoi jäädä koehenkilölle sen verran vieraaksi, että hän arvioi käyttökelpoisiksi sellaisiakin artikkeleita, jotka eivät sitä todellisuudessa olleet. Esimerkiksi hakupyynnön 15 aiheena olivat seurakuntavaalit. Tätä aihepiiriä käsitteleviä artikkeleita ei itse asiassa (myöhemmän analyysin perusteella) löytynyt tulosjoukoista, joten arvioitsijoiden olisi pitänyt todeta, että relevantteja artikkeleita ei ollut ollenkaan. Osa koehenkilöistä kuitenkin arvoi käyttökelpoiseksi kirkkoherranvaalia käsittelevät artikkelit, vaikka seurakunta- ja kirkkoherranvaalit eivät ole sama asia. Ulkopuolisten relevanssiarvioijien käyttämisessä onkin se riski, että relevanttia dokumenttia ei sellaiseksi tunnisteta ja toisaalta kelpuutetaan epärelevantti dokumentti relevantiksi (Harter 1986, s. 165). FULLTEXT-projektissa otettiin se linja, että jatkokäsittelyn pohjana olivat asiantuntijaraadin tekemät relevanssiarviot sellaisinaan. Tutkija ei korjannut arvioita, vaikka olikin muutamien artikkeleiden relevanttiudesta eri mieltä raatinsa kanssa.

Edellämainituista puutteista huolimatta on todettava, että tämän tutkimuksen koejärjestelyt oli toteutettu monin osin paremmin kuin alan klassisissa tutkimuksissa, kuten esimerkiksi Cranfieldissä (vrt. Keen 1992).

7.8 Saannin ja tarkkuuden laskeminen

Tulosjoukkojen saanti- ja tarkkuusarvot laskettiin koehenkilöiden tekemien relevanssiarvioiden perusteella. Näihin laskelmiin otettiin mukaan sekä hy-

vin että jonkin verran relevantit artikkelit, ts. relevanssin astetta ei eroteltu keskiarvolaskuissa.

Kunkin kyselytyypin keskimääräinen tarkkuus ja suhteellinen saanti laskettiin makrokeskiarvon periaatteella. Jokaisen haun tarkkuus- ja saantiarvot laskettiin erikseen ja näistä arvoista otettiin edelleen keskiarvo. (Saannin ja tarkkuuden laskukaavat on selostettu luvussa 3.2).

Saantiarvot laskettiin ns. **suhteellisen** eikä absoluuttisen saannin periaatteella. Jos jollain kyselytyypillä saatiin tulokseksi kaikki saantikannan artikkelit, se sai saantiarvokseen 100 %. Tarkkuus puolestaan ilmaisi tulosjoukon **absoluuttisen** tarkkuuden eli sen, kuinka suuri osuus tulosjoukon dokumenteista oli relevantteja.

7.9 Otantaperiaatteet

Jotta arvioitavien dokumenttien määrä ei FULLTEXT-projektissa olisi paisunut liian suureksi, suurista tulosjoukoista päätettiin ottaa otanta. Suureksi tulosjoukoksi katsottiin tulosjoukko, jossa oli yli 75 dokumenttia.

Esimerkiksi hakupyynnöstä 8 (Kirkkojen rakentaminen) laadittu yhdistelmäkysely ABC tuotti T4-ympäristössä 145 dokumenttia, kun hakusanat yhdistettiin toisiinsa JA-operaattorilla. Kun hakusanat puolestaan yhdistettiin virkeoperaattorilla, yhdistelmäkyselyn tulokseksi saatiin vain 40 dokumenttia. Tässä tapauksessa siis JA-operaattoria käyttäen saadusta 145 dokumentin tulosjoukosta otettaisiin otanta, mutta virkeoperaattorilla saadusta 40 dokumentin tulosjoukosta ei.

Otoksen poiminta tehtiin seuraavan periaatteen mukaan: Otos on aina vakiokokoinen eli 25 dokumenttia. Tämän perusteella systemaattisen otannan otantaväli laskettiin laskukaavalla (6):

$$(6) \qquad \text{floor}\left(\frac{\text{tulosjoukon koko}}{25}\right)$$

Floor(X) pyöristää reaaliluvun X alaspäin seuraavaan kokonaislukuun, jos X ei jo ole kokonaisluku. Jos tulosjoukossa oli esimerkiksi 178 dokumenttia, otantaväli oli floor(178/25) = 7 eli joka seitsemäs dokumentti tuli mukaan otokseen.

Tiettyyn hakupyyntöön eri kyselytyypeillä saadut tulosjoukot yhdistettiin yhdeksi koontijoukoksi (ns. pooling) siten, että alkuperäiset 75 dokumentin ja sitä pienemmät tulosjoukot otettiin sellaisenaan ja suurista tulosjoukoista

taas otettiin otannan tuloksena saadut 25 dokumentin joukot. Kaikkien tulosjoukkojen dokumentit yhdistettiin yhteen nippuun, josta sitten poistettiin kaksoiskappaleet eli dokumentit, jotka esiintyivät useammassa kuin yhdessä tulosjoukossa. Näin kukin dokumentti esiintyi nipussa vain kerran riippumatta siitä, kuinka monen kyselytyypin tulosjoukkoon se muuten oli sisältynyt.

Tulosjoukkojen analyysivaiheessa kuitenkin kävi ilmi, että otosten käyttäminen tuotti ongelmia saanti- ja tarkkuusarvojen vertailuissa. Tämä koski erityisesti kyselyjen laajentamista yhdyssanan osiin eli kyselytyyppejä osien perushausta Aa osien yhdistelmähakuun ABCabc asti.

Ongelmana oli, että eräissä kyselyissä tulosjoukkoon saatiin jokseenkin kaikki relevantit dokumentit käytännössä jo silloin, kun yhdyssanojen osat oli yhdistetty toisiinsa virkeoperaattorilla. Kun kyselyssä sitten käytettiin JA-operaattoria virkeoperaattorin sijasta, näin saadun tulosjoukon koko saattoi olla huomattavan suuri (jopa satoja dokumentteja), mutta joukossa ei juuri ollut uusia relevantteja dokumentteja verrattuna virkeoperaattorilla saatuun tulosjoukkoon. Koska relevanttien dokumenttien osuus näissä JA-operaattoria käyttäen saaduissa tulosjoukoissa oli pieni, relevantit dokumentit jäivät helposti otoksen ulkopuolelle, eli niiden määrä otoksessa oli sattumanvarainen. Joissain vertailuissa kävi niin, että otantana saadun osien yhdistelmäkyselyn (ABCabc) saantiarvo näytti olevan huonompi kuin osien yhdyssanakyselyn (ACac). Tämähän ei voi pitää paikkaansa, koska osien yhdyssanakyselyn tulosjoukko on osien yhdistelmäkyselyn tulosjoukon osajoukko - yhdistelmäkyselyn saannin on loogisesti oltava joko sama tai suurempi kuin sitä suppeammilla kyselytyypeillä saatujen tulosjoukkojen saannin.

Koska tutkimuksen testiympäristö oli jo purettu otosten analyysin aikana, ei enää voitu tehdä toisenlaista otantaratkaisua - kaikkein suurimmista tulosjoukoista oli tulostettu vain otantaan sisältyneet dokumentit. Jos jonkin muun otantamenetelmän tuloksena olisi saatu toisenlainen dokumenttijoukko, ei uusia, pooliin kuulumattomia dokumentteja enää olisi voinut jäljittää eikä niille tehdä relevanssiarvioita.

Koska kaikki peruskyselyn, johdoskyselyn, yhdyssanakyselyn ja yhdistelmäkyselyn tuottamat tulosjoukot olivat käytettävissä kokonaisuudessaan, päätettiin näiden tapauksissa käyttää koko tulosjoukkoa sen koosta riippumatta - otantajoukkoa ei siis käytetty, vaan alkuperäistä tulosjoukkoa, vaikka sen koko olisi ollut suurempi kuin 75 dokumenttia. Tämä ratkaisu oli mahdolli-

nen, koska näiden tulosjoukkojen kaikista dokumenteista olivat myös relevanssiarviot olemassa. Näin näissä vertailuissa ei ole otannasta johtuvia vääristymiä.

Sen sijaan osien peruskyselyn, osien johdoskyselyn, osien yhdyssanakyselyn ja osien yhdistelmäkyselyn tapauksissa ei tulosjoukkojen kaikista dokumenteista ollut tehty relevanssiarvioita - arviot oli tehty vain pienistä tulosjoukoista ja otoksiin sisältyneistä dokumenteista. Tämän vuoksi päädyttiin ratkaisuun, jossa otoksena saadut tulosjoukot jätettiin pois saanti- ja tarkkuusarvoja sekä merkitsevyystestejä laskettaessa. Vertailuihin sisältyneet hakupyynnöt ja niiden tulosjoukot on selostettu tarkemmin luvun 9 alussa.

Summa summarum: Tässä tutkimuksessa otanta ei vaikuttanut saanti- ja tarkkuusarvioihin suoraan, koska otoksena saadut tulosjoukot jätettiin vertailuista pois. Sen sijaan epäsuora vaikutus saantiarvoihin on mahdollinen, sillä saantikannan muodostumiseen otanta vaikutti: saantikantaan sisältyivät myös otannan tuloksena saatujen tulosjoukkojen relevantit artikkelit. Tämä voi periaatteessa laskea tutkimuksessa saatuja saantiarvoja - mitä suurempi saantikanta on, sitä vaikeampi on saavuttaa sadan prosentin saanti.

7.10 Tilastollisen merkitsevyyden laskeminen ja eri vaihtoehtojen välisen eron merkittävyys

Kun vaihtoehtoisia hakujärjestelmiä, hakuympäristöjä tai kyselytyyppejä vertaillaan keskenään testikyselyiden avulla ja eri kyselyillä saatujen tulosjoukkojen todetaan poikkeavan toisistaan, voidaan tilastollisten merkitsevyystestien avulla arvioida, johtuvatko havaitut erot tutkituista muuttujista vai pelkästään sattumasta.

Merkitsevyyden laskeminen aloitetaan määrittelemällä niin sanottu nollahypoteesi H_0 ja sille vaihtoehtoinen hypoteesi H_1. Jos esimerkiksi hypoteesi H_1 väittää, että tietyssä tutkimusympäristöissä saadaan kyselytyyppiä X käytettäessä erilainen tulosjoukko kuin kyselytyyppiä Y käytettäessä, voidaan nollahypoteesinä väittää, että vertailtavien kyselytyyppien X ja Y tuottamat tulosjoukot eivät poikkea toisistaan, eli niiden välillä ei ole eroa. Tilastollisen testin avulla lasketaan, löytyykö nollahypoteesille tukea. Tutkija päättää, mikä on riittävä merkitsevyystaso α eli todennäköisyys, jolla H_0 voidaan hylätä ja todeta vaihtoehtoinen hypoteesi H_1 päteväksi. Yleisesti käytettyjä merkitsevyystasoja ovat esimerkiksi 0.05 ja 0.01. Tasolla 0.05 hyväksytään, että viidessä tapauksessa sadasta hypoteesin H_1 hyväksyminen selittyykin

sattumasta eikä systemaattisesta erosta tulosjoukoissa. (Siegel & Castellan 1988, s. 8 - 9)

Erilaisia tilastotestejä on suuri joukko. Kulloisessakin tutkimuksessa tulisi käyttää testiä, joka sopii tutkittavaan aineistoon. Kinnucanin et al. (1987) katsauksessa kuvataan, miten tilastollisia testejä on sovellettu tiedonhaun tutkimuksessa. He toteavat, että merkitsevyystestejä ei pitäisi tehdä pareittain, kun tutkittavia ryhmiä on enemmän kuin kaksi, koska ryhmien väliset erot saattavat jäädä näkymättömiin pareittaisessa vertailussa. Näissä tapauksissa olisi pareittaisen testin sijasta käytettävä useamman ryhmän vertailuihin tarkoitettua merkitsevyystestiä.

Tämän tutkimuksen tilastotesteissä vertailtiin samanaikaisesti useaa ryhmää. Koska Kristensenin (1992) tutkimuksessa oli todettu Friedmanin merkitsevyystesti (Friedman two-way analysis of variance) tähän tarkoitukseen sopivaksi, samaa testiä käytettiin myös tässä tutkimuksessa. Vertailtavia ryhmiä olivat tietyn tutkimusympäristön eri kyselytyypit tai tietty kyselytyyppi eri tutkimusympäristössä.

Friedmanin testi perustuu oletukselle, että vertailtavat ryhmät tai ainakin niiden mediaanit eli keskimmäiset arvot ovat samat (nollahypoteesi). Jos taas vaihtoehtoinen hypoteesi on tosi, pitää ainakin yhden vertaillun ryhmän mediaanin poiketa muista. Mediaani on keskiarvoa luotettavampi vertailumittari, koska yksittäiset poikkeamat eivät välttämättä muuta sitä lainkaan, kun taas keskiarvoon yksittäinen poikkeava arvo voi vaikuttaa paljonkin.

Kristensen (1992) toteutti Friedmanin merkitsevyystestin Siegelin ja Castellanin (1988) esittämien periaatteiden mukaisesti. Tässä tutkimuksessa on kuitenkin sovellettu Conoverin (1980) esittämää laskutapaa, joka poikkeaa Siegelin ja Castellanin esittämästä menetelmästä. Conoverin menetelmä valittiin, koska Hull (1993; 1996) on soveltanut sitä omissa tutkimuksissaan. Hullin mukaan Conoverin esittämä menetelmä ei ole sama kuin alkuperäinen Friedmanin testi, vaan sitä tehokkaampi. Lisäksi Conoverin menetelmä on yksinkertaisempi laskea silloin, kun vertailtavassa aineistossa on paljon sidoksia, kuten FULLTEXT-projektissa oli laita. Sidokset tarkoittavat merkitsevyystestissä tilannetta, jossa vertailtavien vaihtoehtojen arvot ovat täysin samat (eli niiden välillä ei ole minkäänlaista eroa).

Friedmanin merkitystesti laskettiin Conoverin (1980) mukaan seuraavasti (esimerkki yhdestä laskelmasta liitteessä 13 ja yhteenveto merkitsevyyslaskelmien tuloksista liitteessä 14):

1. Vertailtavien ryhmien saanti- tai tarkkuusarvot asetettiin taulukkoon, jossa oli N riviä ja k saraketta. Tässä tutkimuksessa N = hakupyyntöjen määrä ja k = kyselytyyppien tai tutkimusympäristöjen määrä.

2. Kukin taulukossa oleva lukuarvo korvattiin rangillaan eli järjestysluvullaan. Tällöin kullakin rivillä olevat arvot muutettiin järjestysluvuiksi $R(X_{ij})$ yhdestä k:hon niin, että pienin arvo sai järjestysluvun 1; j = 1, 2,…, k ja i = 1, 2,…, N. Jos samalta riviltä löytyi sidoksia eli tapauksia, jossa vertailtujen vaihtoehtojen arvot olivat samat, näille tapauksille annettiin järjestyslukujen keskiarvo.

   Järjestyslukujen keskiarvo lasketaan seuraavasti: Jos rivillä oli esimerkiksi viisi arvoa, joista pienin oli jo saanut järjestysluvun 1 ja kolmella seuraavalla tapauksella oli sama (tässä tapauksessa toiseksi pienin) lukuarvo, annettiin näille kolmelle tapaukselle sama järjestysluku. Tämä järjestysluku laskettiin ottamalla jo annettua järjestyslukua (tässä tapauksessa siis ykköstä) järjestyksessä seuraavien kolmen järjestysluvun keskiarvo, tässä tapauksessa (2+3+4)/3 = 3.

3. Kun järjestysluvut oli laskettu kaikkien rivien kaikille arvoille, laskettiin järjestyslukujen summa R_j sarakkeittain, j = 1, 2,…, k:

   (7) $$R_j = \sum_{i=1}^{N} R(X_{ij})$$

4. Tämän jälkeen laskettiin taulukon kaikkien järjestyslukujen neliöiden summa A kaavalla (8). Mikäli taulukossa ei ole lainkaan sidoksia, voidaan käyttää yksinkertaisempaa laskukaavaa - FULLTEXT-aineistossa ei tällaisia tapauksia kuitenkaan ollut, vaan kaikissa taulukoissa oli sidoksia.

   (8) $$A = \sum_{i=1}^{N}\sum_{j=1}^{k}\left[R(X_{ij})\right]^2$$

5. Seuraavaksi laskettiin arvo B, joka on sarakkeiden j-lukusummien neliöiden keskiarvo. Laskettiin siis sarakkeiden järjestyslukujen summien neliöt yhteen ja jaettiin näin saatu luku hakupyyntöjen määrällä eli luvulla N seuraavasti (R_j = j:nnen sarakkeen järjestyslukujen summa):

   (9) $$B = \frac{1}{N}\sum_{j=1}^{k} R_j^2$$

6. Varsinainen merkitsevyystesti laskettiin seuraavalla kaavalla:

$$T = \frac{(N-1)\left[B - Nk(k+1)^2/4\right]}{A-B} \quad (10)$$

7. Merkitystestin tulosta T verrattiin Conoverin kirjassa esitettyihin taulukoihin (Conover 1980, liitteen taulukko A26). Jos testitulos T oli valitulla merkitsevyystasolla suurempi kuin taulukon arvo, voitiin hylätä nollahypoteesi ja todeta vertailtavien ryhmien eron olevan tilastollisesti merkitsevä.

8. Jos Friedmanin merkitystestillä saatavan testituloksen arvo on suurempi kuin taulukon arvo, tiedetään, että ainakin yksi ryhmistä eroaa vähintään yhdestä toisesta ryhmästä. Se ei kuitenkaan kerro mikä tai mitkä ryhmistä poikkeavat muista. Tämä selvitetään erillisellä jatkotestillä, johon käytetään kaavaa (11). Sen tuloksen perusteella eri ryhmiä verrataan pareittain toisiinsa.

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

$\left|R_i - R_j\right|$ = vertailtavien ryhmien järjestyslukujen summien erotus (erotuksen itseisarvo)

t = kriittinen arvo, kun α on merkitsevyystaso

k = sarakkeiden (kyselytyyppien tai tutkimusympäristöjen) määrä

N = rivien (eli kyselyjen) määrä

Kriittinen arvo $t_{1-\alpha/2}$ saadaan Conoverin kirjan liitteen taulukosta A25. Vapausasteet lasketaan kaavalla $(N-1)(k-1)$.

Jos tutkimusympäristöjen X ja Y tehokkuusmittarina käytetään saanti- ja tarkkuusarvoja, niin kuinka paljon X:n ja Y:n saanti- tai tarkkuusarvojen sitten pitää erota toisistaan, että X tai Y voidaan todeta paremmaksi? Eri lukuarvojen välinen ero voi olla melko pieni, vaikka se todettaisiinkin tilastollisesti merkitseväksi.

Sparck Jones (1974) on määrittellyt ns. **käytännössä merkitsevän eron** (practical significance/importance). Alun pitäen Sparck Jones esitti tämän peukalosäännön automaattiseen indeksointiin tarkoitettujen järjestelmien vertailuja varten: Jos kahden vertailtavan vaihtoehdon välinen ero on tilastollisesti merkitsevä, mutta saannin tai tarkkuuden lukuarvot eroavat toisis-

taan vähemmän kuin 5 prosenttiyksikköä, ero on niin vähäinen, että sitä ei käytännössä kannata ottaa huomioon. Kun vertailtavien vaihtoehtojen ero on tilastollisesti merkitsevä ja suorituskyvyssä (saanti- tai tarkkuusarvoissa) on 5 - 10 prosenttiyksikön ero, eroa suorituskyvyssä sanotaan **huomattavaksi** (noticeable) ja yli 10 prosenttiyksikön eroa **olennaiseksi** (material).

Toisaalta Keen (1992) toteaa, että vaikka vertailtavien muuttujien väliset erot eivät olisi erityisen suuret ja vaikka eroille ei löytyisi tilastollista merkitsevyyttäkään, on oma arvonsa silläkin, jos erot ovat konsistentteja eli yhdenmukaisia (consistency). Esimerkiksi Harmanin (1991) tutkimuksessa päätteiden karsiminen tuotti lähes poikkeuksetta paremman tai vähintäänkin saman tuloksen kuin karsimatta jättäminen. Harmanin itsensä mielestä karsimisen ja karsimatta jättämisen välille ei voinut löytää riittävästi eroja, kun taas Keenin mielestä tulosten yhdenmukaisuus tulisi ottaa huomioon - eli että karsinta enimmäkseen tuottaa parempia tuloksia kuin karsimatta jättäminen.

8 PERUSMUOTOISTAMISEN VAIKUTUS HAKEMISTOSANOIHIN

8.1 Tallennettavan tekstin käsittely

Testitietokannat tuotettiin Aamulehden toimituksellisesta tekstistä ajanjaksolta 2.1. - 31.3.1990. Tietokannat rakennettiin syöttämällä järjestelmään kerrallaan yhden päivän artikkeliaineisto. Päiviä oli yhteensä 89 ja yhden päivän lehti sisälsi keskimäärin 261 artikkelia. Pienin päivittäinen artikkelimäärä oli 151 ja suurin 376 kappaletta. Artikkelien lukumäärä oli tutkimusotoksen lehdissä suurempi kuin Aamulehden arkistojärjestelmässä keskimäärin (Ylinen 1991). Kaikkiaan artikkeleita eli tekstidokumentteja kertyi tutkimustietokantaan 23 244 kappaletta.

Dokumenttien eri kentät sisälsivät merkkijonoja yhteensä 3 652 073 kappaletta. Perusmuotohakemistoja tuotettaessa ainoastaan otsikko- ja tekstikenttien sisältö syötettiin suomen kielen tulkintaohjelmalle; muut kentät sisälsivät julkaisun yksilöintitietoja, kuten kirjoittajan nimen tai päivämäärän, joille ei ollut järkevää tehdä morfologista analyysiä (taulukko 3). Otsikko- ja tekstikentät kuitenkin sisälsivät suurimman osan tietueen merkkijonoista, 3 582 356 kappaletta. Tämä oli keskimäärin 98,1 % käsiteltävien merkkijonojen kokonaismäärästä.

Taulukko 3. BASIS-hakujärjestelmään tallennettujen tekstidokumenttien sisältämä merkkijonojen määrä sekä otsikko- ja tekstikenttien sisältämien sananmuotojen osuus näistä. Lisäksi "Päivittäin" -kohdassa on päivittäisten syötteiden vertailun tuloksia.

	Dokum. kaikki merkki-jonot, kpl	Tekstin ja otsikon merkki-jonot, kpl	Tekstin ja otsikon mj:t, %-osuus	Muiden kenttien merkki-jonot, kpl	Muiden kenttien mj:t, %-osuus
Yhteensä	3 652 073	3 582 356	98,1 %	69 717	1,9 %
Päivittäin:					
-keskiarvo	41 035	40 251	98,1 %	783	1,9 %
-mediaani	39 997	39 140	98,1 %	763	1,9 %
-pienimmillään	26 160	25 596	97,6 %	453	1,5 %
-suurimmillaan	59 431	58 408	98,5 %	1 128	2,4 %

Yksittäisen päivän artikkeliaineistossa eli yhden päivän syötteessä oli keskimäärin 41 035 merkkijonoa. Alimmillaan yhden päivän syötteessä oli yhteensä 26 160 merkkijonoa. Suurimmassa yhden päivän syöteaineistossa oli yhteensä 59 431 merkkijonoa. Päivittäisistä syötteistä laskien otsikko- ja tekstikenttien merkkijonojen osuus tallennetuista merkkijonoista oli pienimmillään 97,6 % (jolloin muiden kenttien merkkijonojen osuus oli 2,4 %) ja suurimmillaan 98,5 % (muiden kenttien merkkijonojen osuus 1,5 %).

Hakemistoon tallennettavien sanojen perusmuotoistamista testattiin pääasiassa Morfo-ohjelmalla. Twol-ohjelmalla tehtiin suppea testiajo, jossa lähinnä testattiin sen toimivuutta BASIS-hakujärjestelmässä (luku 8.5).

Kun Morfo pelkästään palautti sananmuodot perusmuotoon, otsikko- ja tekstikenttien sisältämistä 3 582 356 sananmuodosta syntyi 4 155 225 perusmuotoa; merkkijonojen kokonaismäärä siis kasvoi perusmuotoistamisen seurauksena 16 %. Kasvu johtuu monitulkintaisista eli homografisista sananmuodoista: taivutusmuotohakemistoon otsikko- ja tekstikenttien sananmuodot tallennetaan sellaisinaan, mutta perusmuoto-ohjelmat löytävät homografeille useamman kuin yhden perusmuototulkinnan, jotka kaikki on tallennettava hakemistoon. (Taulukko 4.)

Kun Morfo perusmuotoistamisen lisäksi jakoi yhdyssanat osiinsa ja tuotti näistä osista vielä kaikki mahdolliset yhdistelmät, perusmuotoja ja yhdyssanan osina esiintyviä sanoja syntyi yhteensä 5 292 582 kappaletta. Tämä määrä on 147,7 % syötettyjen sananmuotojen määrästä. Tässä tapauksessa

Taulukko 4. Morfo-ohjelman tuottamien perusmuotojen, perusmuotojen ja yhdyssanan osien sekä tunnistamatta jääneiden sananmuotojen lukumäärä BASIS-hakujärjestelmässä sekä niiden suhde käsiteltyjen sananmuotojen kokonaismäärään prosentein ilmaistuna. Lisäksi "Päivittäin" -kohdassa on päivittäisten syötteiden vertailun tuloksia.

	Vain perusmuodot, kpl	Vain perusmuodot, %	Perusm. + osat, kpl	Perusm. + osat, %	Tunnistamatta, kpl	Tunnistamatta, %
Yhteensä	4 155 225	116,0 %	5 292 582	147,7 %	265 711	7,4 %
Päivittäin:						
-keskiarvo	46 688	116,0 %	59 467	147,8 %	2 986	7,4 %
-minimi	30 122	106,9 %	38 231	133,7 %	1 126	4,1 %
-maksimi	66 732	122,2 %	85 700	159,2 %	6 024	12,4 %

47,7 %:n lisäyksen aiheuttivat sekä monitulkintaiset sananmuodot että yhdyssanojen osat. (Taulukko 4.)

Perusmuotoistamisessa syntyvien tulkintojen määrä riippuu perusmuotoohjelman sanakirjasta: mitä enemmän siinä on sanoja, sitä suurempi osuus sananmuodoista saadaan tunnistetuksi. Toisaalta ohjelma samalla löytää suhteellisesti enemmän myös monitulkintaisia sananmuotoja. Pelkkiä perusmuotoja tuotettaessa Morfo oli tarkoituksellisesti säädetty ylitulkitsevaksi (lähinnä yhdyssanojen suhteen), joten merkkijonojen määrän suuri lisäys oli odotettavissakin. Mikäli Morfo-ohjelman tulkintatapa määriteltäisiin tarkemmaksi kuin tässä tutkimuksessa ja lisäksi käytettävissä olisi monitulkintaisuuksia karsiva disambiguointiohjelma, erilaisia perusmuotoja kertyisi vähemmän.

Morfo-ohjelmalta jäi 3 582 356 sananmuodosta tunnistamatta 265 711 kappaletta, keskimäärin 7,4 % syötteestä. Nämä merkkijonot tallennettiin sellaisinaan tunnistamattomien sananmuotojen hakemistoon. Kun eri päivien syötteitä ja perusmuotoistamisen tuloksia verrataan keskenään, tunnistamatta jääneiden sananmuotojen osuus oli suurimmillaan 12,4 % ja pienimmillään 4,1 % (taulukko 4).

Tunnistamatta jääneitä ilmauksia ei systemaattisesti analysoitu. Yleisvaikutelmaksi jäi, että lehtitekstistä jäivät tunnistamatta lähinnä kirjoitusvirheen sisältävät sananmuodot sekä vierasperäiset ja suomalaiset erisnimet. Joukossa oli myös yleisnimiä, joiden periaatteessa olisi pitänyt sisältyä Morfo-ohjelman sanakirjaan. Tällaisia kuitenkin esiintyi tunnistamattomien sanojen hakemistossa huomattavasti harvemmin kuin edellämainittuja muita ongelmailmauksia. Vertailun vuoksi voidaan mainita, että kun englanninkielisiä talousuutisia analysoitiin Raun (1991) projektissa, tekstin sananmuodoista jäi tunnistamatta yli 8 prosenttia. Talousuutisten kaikista sanoista oli yritysten nimien osuus yli 4 prosenttia; näistä tunnistamatta jäi neljännes.

8.2 Merkkijonojen määrä hakemistossa

Morfo-ohjelman tuottamien merkkijonojen määrästä voisi ensisilmäyksellä luulla, että, perusmuotohakemistoihin (H2, H3) tulee enemmän hakemistosanoja kuin taivutusmuotohakemistoon (H1). Näin ei kuitenkaan todellisuudessa tapahdu. Vaikka merkkijonoja onkin perusmuotohakemistoissa **luku-**

määräisesti enemmän, niiden joukossa on **vähemmän erilaisia** merkkijonoja kuin taivutusmuotohakemistossa, jossa eri taivutusmuodot tallennetaan erikseen omina merkkijonoinaan.

Merkkijonojen lukumäärät laskettiin kolmesta hakemistosta (H1, H2 ja H3) jokaisen päivitysajon jälkeen eli kun yhden päivän artikkelit oli syötetty tietokantaan. Lukumäärissä olivat mukana kaikkien kenttien sisältämät merkkijonot, siis myös muut kuin otsikko- ja tekstikenttien sananmuodot. Koska kaksoishakemisto (H1+3) tuotettiin eri tavalla kuin nämä kolme muuta hakemistoa, siitä ei mitattu hakemistosanojen ja osoitteiden määrän päivittäistä kasvua.

Eniten erilaisia merkkijonoja oli taivutusmuotohakemistossa (H1). Kun taivutusmuodot pelkästään palautettiin perusmuotoon ja tallennettiin nämä perusmuodot hakemistoon (sekä tunnistamattomat sananmuodot omaan hakemistoonsa), merkkijonojen kokonaismäärä H2-hakemistossa oli vain noin puolet taivutusmuotohakemiston merkkijonojen määrästä. Kun perusmuotoistamisen lisäksi jaettiin yhdyssanat osiinsa ja osat sekä niiden yhdistelmät (sekä tunnistamatta jääneet sananmuodot) tallennettiin hakemistoon (H3), merkkijonojen lukumäärä tässä hakemistossa oli noin 60 % taivutusmuotohakemiston merkkijonojen määrästä. (Taulukko 5.)

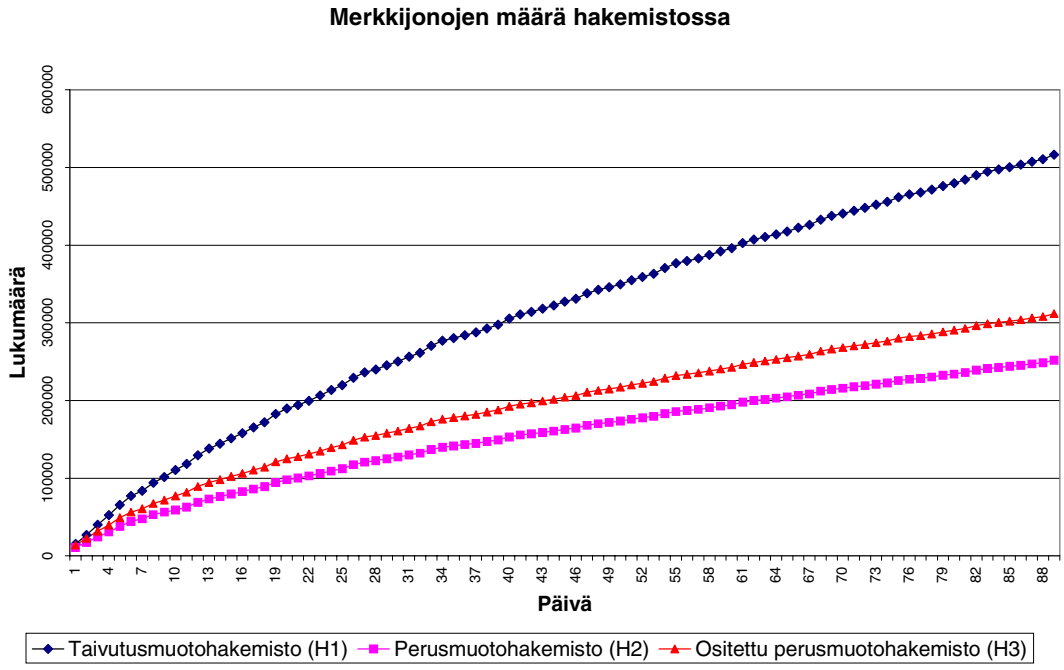
Tutkimustulokset siis tukevat hypoteesia 1, jonka mukaan perusmuotoistaminen tallennusvaiheessa vähentää hakemistoon tulevien merkkijonojen määrää. Perusmuotohakemistoihin ei tarvitse jatkuvasti lisätä uusia merkkijonoja, vaan niissä selvitään taivutusmuotohakemistoa useammin sillä, että hakemistossa jo olevaan sanaan liitetään uusi osoite. Kuvassa 11 näkyvät merkkijonojen päivittäiset lisäykset ja merkkijonojen lukumäärän kasvu eri hakemistoissa. Vaikka yhdyssanat jaettaisiinkin osiinsa, jää merkkijonojen

Taulukko 5. Merkkijonojen määrä eri hakemistossa sekä niiden määrän suhde taivutusmuotohakemiston merkkijonojen määrään.

	Taivutus-muotohake-misto (H1)	Perusmuoto-hakemisto (H2)	Ositettu perusmuoto-hakem. (H3)
Merkkijonojen lukumäärä	516 371	251 763	311 707
Suhteessa taivutus-muotohakemiston merkkijonojen määrään	100,0 %	48,8 %	60,4 %

määrä myös ositetussa perusmuotohakemistossa selvästi pienemmäksi kuin taivutusmuotohakemistossa.

Taulukossa 5 näkyvät tulokset ovat samansuuntaisia kuin ne, jotka Niemistö (1988, s. 39) sai vastaavantyyppisissä hakemistoissa. (FULLTEXT-projektissa ja Niemistön tutkimuksessa käytetyt hakemistot poikkesivat joiltain osin toisistaan, esimerkiksi yhdyssanojen ja sulkusanalistan käsittelyn suhteen.) Niemistön tutkimuksessa hakemistosanoja oli perusmuotoon palauttamisen jälkeen 50 % vähemmän kuin taivutusmuotohakemistossa sananmuotoja eli tulos oli suunnilleen sama kuin tässä tutkimuksessa. Kun perusmuotoon palauttamisen lisäksi oli hajoitettu yhdyssanat osiinsa ja tehty osista kaikki mahdolliset yhdistelmät, Niemistön testaamassa hakemistossa hakemistosanojen määrä oli 55 % taivutusmuotohakemiston sananmuotojen määrästä; tässä tutkimuksessa vastaava luku oli 60 %.



Kuva 11. Merkkijonojen määrän päivittäinen kasvu BASIS-hakemistoissa. Erilaisia merkkijonoja oli eniten taivutusmuotohakemistossa; vähiten erilaisia merkkijonoja oli hakemistossa, jossa sananmuodot oli pelkästään palautettu perusmuotoon.

8.3 Osoitteiden määrä hakemistossa

Vähiten osoitteita kertyi taivutusmuotohakemistoon (H1). Pelkät perusmuodot (ja tunnistamattomat sananmuodot) sisältävässä hakemistossa (H2) osoitteita oli yli 9 % enemmän kuin taivutusmuotohakemistossa. Kasvu joh-

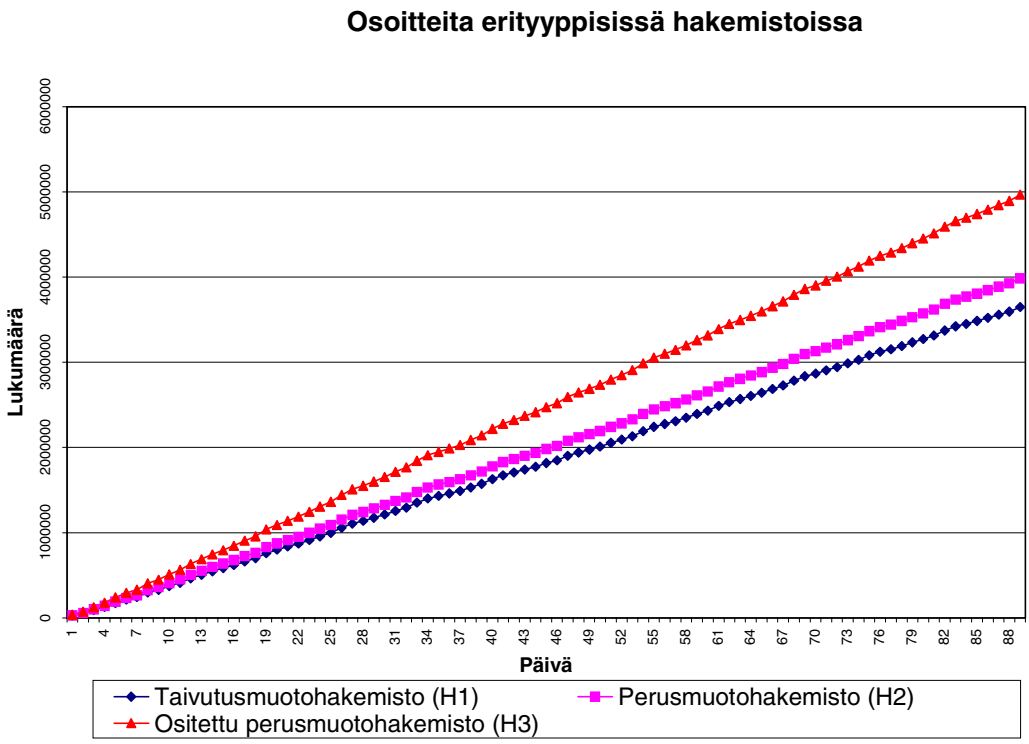
tui monitulkintaisista sanoista: kun sananmuoto voidaan tulkita usealla eri tavalla eli siitä voidaan tuottaa useampi perusmuoto, on näihin jokaiseen mahdolliseen tulkintaan (hakemistosanaan) liitettävä alkuperäisen sananmuodon osoitetiedot.

Perusmuodot sekä yhdyssanojen osat yhdistelmineen sisältävä hakemisto (H3) puolestaan sisälsi osoitteita yli 36 % enemmän kuin taivutusmuotohakemisto. Tässä lisäys johtui sekä monitulkintaisista sanoista että siitä, että yhdyssanan osoite oli liitettävä sanan perusmuodon lisäksi myös sen kaikkiin osiin. (Taulukko 6 ja kuva 12; luvuissa ovat mukana kaikki kentät, eivät vain otsikko- ja tekstikentät.)

Osoitteiden lukumääriä tarkasteltaessa on otettava huomioon, että osoitteet merkittiin virkkeen tarkkuudella. Perussääntö on, että mitä tarkempia osoitteet ovat, sitä suurempi määrä niitä on dokumenttia kohden. Suurin määrä osoitteita syntyy, kun osoitteet on ilmaistu sanan/merkkijonon tarkkuudella, ts. kun osoite ilmaisee tietyn merkkijonon tarkan sijainnin dokumentin sisällä (esimerkiksi *n*:s merkkijono dokumentin alusta). Kun osoite ilmaistaan virkkeen tarkkuudella, tiedetään, missä virkkeessä sana on, muttei sen tarkempaa sijaintia virkkeen sisällä. Koska jokin sana(nmuoto) saattaa esiintyä virkkeessä useammin kuin kerran, päällekkäiset osoitetiedot (duplikaatit) voidaan jättää hakemistosta pois. Edelleen, kun osoite ilmoitetaan vieläkin yleisemmin eli kappaleen tai dokumentin tarkkuudella, on yhä todennäköisempää, että sama sana(nmuoto) esiintyy siinä useammin kuin kerran, jolloin poisjätettäviä päällekkäisiä osoitteita on vielä enemmän. Toisaalta tietokannan tuottajan on osoitteiden tarkkuustasoa päättäessään otettava huomioon muutkin näkökohdat kuin muistitilan kulutus; hakuvaiheessa liian epätarkat osoitteet tuottavat ongelmia, mikäli käyttäjä ei pysty kohdistamaan hakua riittävän tarkalle tasolle.

Taulukko 6. Osoitteiden kokonaismäärä perusmuotohakemistossa ja ositetussa perusmuotohakemistossa sekä osoitteiden määrän suhde taivutusmuotohakemiston osoitteiden määrään.

	Taivutus-muotohake-misto (H1)	Perusmuoto-hakemisto (H2)	Ositettu perusm. hak. (H3)
Osoitteiden lukumäärä	3 648 784	3 985 311	4 966 492
Suhteessa taivutus-muotohakemiston osoitteiden määrään	100,0 %	109,2 %	136,1 %



Kuva 12. Osoitteiden määrän päivittäinen kasvu eri hakemistoissa. Osoitteita oli vähiten taivutusmuotohakemistossa.

Niemistön tutkimuksessa osoitteet merkittiin eri tarkkuustasolla kuin tämän tutkimuksen BASIS-hakemistoissa. Niemistöllä osoitteet ilmaisivat dokumentin, jossa sana oli esiintynyt, sekä sanan tarkan sijainnin ko. dokumentissa (Niemistö 1988, s. 75). Mutta vaikka FULLTEXT-projektin ja Niemistön tulokset mitattiin erilaisista hakemistoista, ne ovat hyvin lähellä toisiaan: Niemistöllä osoitteiden määrä oli perusmuotohakemistossa noin 11 % suurempi kuin taivutusmuotohakemistossa, kun se tässä tutkimuksessa oli noin 9 % suurempi. Kun Niemistön tutkimuksessa perusmuotoistamisen lisäksi jaettiin yhdyssanat osiin ja osista tuotettiin kaikki yhdistelmät, osoitteiden määrä kasvoi taivutusmuotohakemistoon verrattuna 36 % eli saman verran kuin tässä tutkimuksessa.

8.4 Hakemiston muistitila

Sananmuotojen palauttaminen perusmuotoon siis vaikuttaa kahtalaisesti: se vähentää erilaisten merkkijonojen lukumäärää hakemistossa, mutta lisää osoitteiden määrää. Näiden kahden vastakkaisen suuntauksen kokonaisvaikutus nähdään laskemalla hakemistojen muistitila.

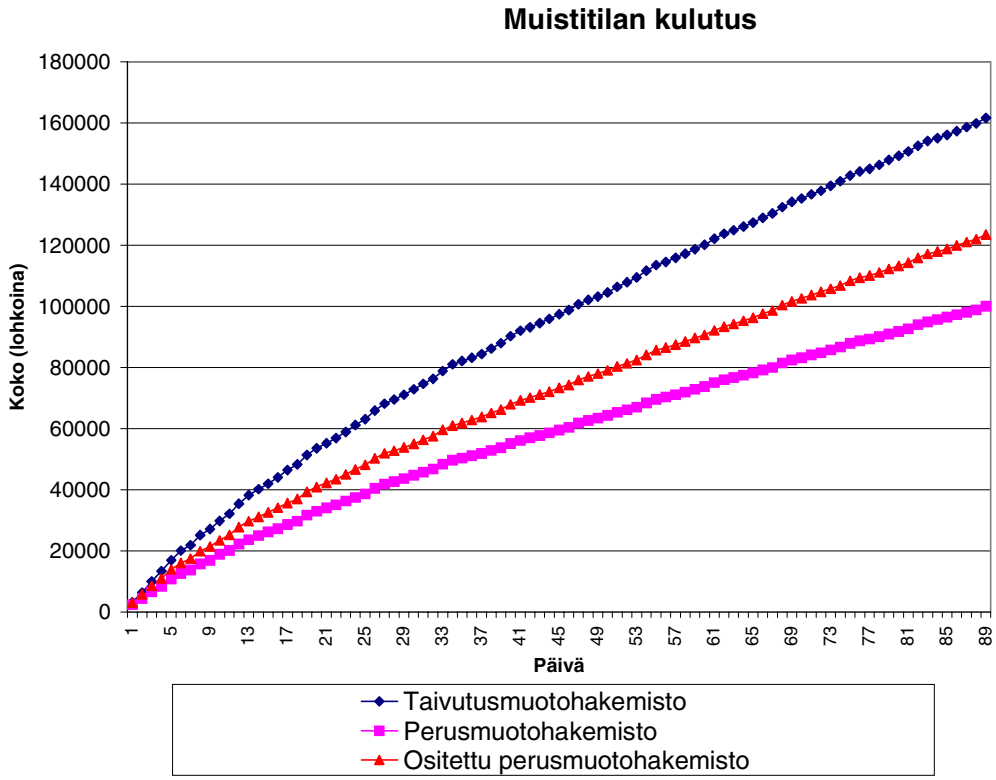
BASIS-hakujärjestelmän tuottamista seurantatiedoista nähtiin, kuinka monta lohkoa eri hakemistot kuluttivat muistitilaa. (Yksi lohko varaa VAX-laitteiston muistitilaa 512 merkkiä eli tavua.) Koska muistitila mitattiin vasta lopputilanteessa eli kun kaikki artikkelit oli syötetty testitietokantoihin, kaikki neljä testihakemistoa (H1, H2, H3 ja H1+3) olivat rinnastettavissa keskenään (taulukko 7). Tosin tulosten tulkinnassa on muistettava, että hakemisto H1+3 tuotettiin hakemistosta H3 ja sen muistitilan käyttö oli optimoitu H3-hakemiston mukaisesti. Todennäköisesti H1+3-hakemistolle olisi voitu kokeilemalla löytää jokin optimaalisempi jako fyysisiin segmentteihin, jolloin muistitilaa olisi kulunut jonkin verran vähemmän (ks. luku 7.3.4).

Vertailujen tuloksena todettiin, että pelkät perusmuodot sekä tunnistamattomat sananmuodot sisältävä hakemisto (H2) oli kooltaan 62 % taivutusmuotohakemistosta (H1). Kun hakemistossa olivat perusmuotojen lisäksi myös yhdyssanat sekä niiden osat kaikkine yhdistelmineen (H3), se oli kooltaan 76 % taivutusmuotohakemistosta eli tämäkin hakemisto oli taivutusmuotohakemistoa pienempi. Kun tietokanta oli toteutettu niin, että taivutusmuotohakemisto ja ositettu perusmuotohakemisto olivat rinnakkain (hakemisto H1+3), tällaisen kaksoishakemiston muistitilan tarve luonnollisestikin oli suurempi kuin pelkän taivutusmuotohakemiston. Tässä tutkimuksessa toteutetun kaksoishakemiston tilantarve oli 62 % suurempi kuin taivutusmuotohakemiston yksinään. Hakemiston toteutustapa ei kuitenkaan ollut optimaalinen, joten normaalioloissa tilantarve todennäköisesti jäisi tätä pienemmäksi.

Taulukko 7. Eri hakemistojen muistitila sekä niiden suhde taivutusmuotohakemiston muistitilaan (määrä ilmaistu sekä 512 merkin lohkoina että megatavuina).

	Taivutus-muotohake-misto (H1)	Perus-muotoh. (H2)	Ositettu perusm. hak. (H3)	Kaksois-hakemisto (H1+3)
Muistitilan määrä lohkoina	161 630	100 070	123 520	261 788
Muistitilan määrä megatavuina (Mt)	82,75	51,24	63,24	134,03
%-osuus taivutus-muotohakemistosta	100,0 %	61,9 %	76,4 %	162,0 %

Hakemistojen H1, H2 ja H3 muistitilan päivittäinen kehitys näkyy kuvasta 13. Vertailujen lopputuloksena siis oli, että merkkijonojen lukumäärän väheneneminen pienentää hakemiston kokoa enemmän kuin osoitteiden määrän lisääntyminen sitä kasvattaa.



Kuva 13. Hakemistojen H1, H2 ja H3 muistitilojen päivittäinen kasvu.

Hakemiston koko riippuu sekä kielestä, jolla tallennettavat dokumentit on kirjoitettu, että tavasta, jolla hakemisto on toteutettu. Nyrkkisääntönä voidaan pitää, että englanninkielisiä tekstejä sisältävissä tietokannoissa hakemiston ja varsinaisen dokumenttitiedoston suhde on normaalitapauksissa 1:1. Tällöin hakemisto on siis yhtä suuri kuin peräkkäistiedosto, johon itse tekstidokumentit on tallennettu (Salton 1989, s. 206).

Mikäli hakemisto rakennetaan normaalitapaa tehokkaammin, voidaan hakemiston ja varsinaisen dokumenttitiedoston suhde saada edullisemmaksi. Joissain hakujärjestelmissä hakemiston koon sanotaan olevan vain viidesosan dokumenttitiedoston koosta (Newton 1983). Zobel et al. (1995) puolestaan mainitsevat, että heidän soveltamillaan tiivistysmenetelmillä hakemiston tilantarve saadaan kutistettua niin, että se vie vain 10 % alkuperäis-

Taulukko 8. FULLTEXT-projektin eri hakemistojen suhde dokumenttitiedostoon eli peräkkäistiedostoon, joka sisältää varsinaiset tekstidokumentit.

Muistitila	Dokumenttitiedosto	Taivutusmuotoh. (H1)	Perusmuotoh. (H2)	Ositettu perusm. (H3)	Kaksoishakem. (H1+3)
Tiedoston koko lohkoina	173 208	161 630	100 070	123 520	261 788
Tiedoston koko megatavuina (Mt)	88,68	82,75	51,24	63,24	134,03
Hakemiston suhde dok.tiedostoon, %		93,3 %	57,8 %	71,3 %	151,1 %

tekstin tarvitsemasta muistitilasta (kun osoitteet on ilmaistu dokumentin tasolla).

Muissa kielissä hakemiston ja dokumenttitiedoston suhdetta ei välttämättä saada yhtä edulliseksi kuin englannin kielessä. Kielen ominaisuuksien ja hakemiston tehottoman toteutustavan seurauksena muistitilan tarve voi moninkertaistua. Esimerkiksi eräässä hakujärjestelmässä hakemiston suhde dokumenttitiedostoon oli 1-2:1 englanninkielistä aineistoa tallennettaessa, kun taas suomenkielistä aineistoa käsiteltäessä päästiin parhaimmillaankin vain suhteeseen 5:1 (Niemistö 1988, s. 14)[1]. Samaten saksankielisestä aineistosta tuotettu hakemisto on kooltaan suurempi kuin vastaava englanninkielinen hakemisto (Kotzias 1990).

Taivutusmuotohakemiston (H1) koko oli suomenkieliseksi hakemistoksi varsin edullinen, hiukan yli 90 % varsinaisen dokumenttitiedoston koosta (taulukko 8). Hakemistoon tallennettavien sananmuotojen perusmuotoistaminen kuitenkin muutti hakemiston (H2) ja peräkkäistiedoston suhteen edullisemmaksi, sillä hakemiston koko oli tällöin vajaat 60 % dokumenttitiedoston koosta. Vielä silloinkin, kun perusmuotohakemisto sisälsi myös yhdyssanojen osat ja niiden kaikki yhdistelmät (H3), se oli pienempi kuin taivutusmuotohakemisto eli noin 70 % peräkkäistiedoston koosta.

[1] Ilmeisesti kyseinen järjestelmä varaa erityisesti hakemiston merkkijonoille paljon tilaa: mitä enemmän hakemistoon tulee erilaisia merkkijonoja, sitä enemmän tilaa tarvitaan. Tässä tutkimuksessa käytetyllä BASIS-järjestelmällä hakemiston ja peräkkäistiedoston välinen suhde oli kaikissa kokeilluissa vaihtoehdoissa huomattavasti parempi.

8.5 Twol-ohjelmiston testaus

Lingsoft Oy:n Twol-ohjelmisto oli käytettävissä vain FULLTEXT-projektin loppuajan, joten sillä ei tehty yhtä laajoja testauksia kuin Kielikone Oy:n Morfo-ohjelmistolla. Yleiskuvan saamiseksi Twol- ja Morfo-ohjelmien tuottamia analyysituloksia verrattiin BASIS-hakujärjestelmässä siten, että molempien tulkintaohjelmien syötteenä oli yhden ja saman päivän aineisto. Vertailussa käytettiin hakemistoja, jossa sananmuodot vain palautettiin perusmuotoonsa (H2); vaihtoehtoa, jossa myös yhdyssanat olisi jaettu osiinsa, ei toteutettu. Taulukossa 9 ovat tiedot Twol-ohjelmalle annetusta syötteestä ja taulukossa 10 on vertailtu Twol- ja Morfo-ohjelmien tuottamia analyysituloksia keskenään.

Sananmuodon tulkinta (tai tunnistamatta jääminen) riippuu perusmuoto-ohjelman sanakirjan kattavuudesta. Morfo-ohjelman sanakirjassa oli tutkimuksen ajankohtana noin 60 000 erilaista perusmuotoa, joiden lisäksi se pystyi tunnistamaan näitä perussanoja sisältävät yhdyssanat. Twol-ohjelman suomenkielinen sanakirja sisälsi vastaavasti noin 37 000 perusmuotoa (Karetnyk et al. 1991). Mitä enemmän sanakirjassa on sanoja, sitä suurempi osuus sananmuodoista tunnistetaan, mikä tietenkin on positiivista. Haittana kuitenkin on, että samalla yhä useammalle sanalle löydetään enemmän kuin yksi tulkinta. Kun ei tiedetä, mikä vaihtoehdoista on oikea, on ne kaikki tallennettava hakemistoon. Vaikka ylimääräisten tulkintojen tallentaminen hakemistoon ei kovin paljon lisäisikään muistitilan kulutusta, jokainen niistä on käytännössä väärä osoite. Hakuvaiheessa ne tuottavat tulosjoukkoon dokumentteja, jotka eivät todellisuudessa sisälläkään hakusanaa. Sanakirjaa ei siis kannata kasvattaa miten suureksi tahansa, vaan olisi löydettävä sellainen sanakirjan koko, jolla sekä tunnistamattomien että monitulkintaisten sananmuotojen osuus jää mahdollisimman pieneksi.

Taulukko 9. Vertailuaineistona olleeseen yhden päivän artikkelijoukkoon sisältyneiden merkkijonojen määrä sekä otsikko- ja tekstikenttien sisältämien sananmuotojen osuus näistä.

Artikkelien merkki-jonot, kpl	Teksti ja otsikko, kpl (Twol:lle)	Teksti ja otsikko, %-osuus	Muut, kpl (ei syötetty Twol:lle)	Muut, %-osuus
27 747	27 174	97,9 %	573	2,1 %

Taulukko 10. Sananmuotoanalyysin tuloksena saatujen perusmuotojen ja tunnistamatta jääneiden sananmuotojen lukumäärä sekä niiden osuus suhteessa käsiteltyjen sananmuotojen määrään (27 147 kpl).

	Perus-muodot, kpl	Perus-muodot, %	Tunnis-tamatta, kpl	Tunnis-tamatta, %
Twol	27 502	101,2 %	2 641	9,7 %
Morfo	30 908	113,7 %	2 315	8,5 %

Vertailujen perusteella Twol- ja Morfo-ohjelmien välillä ei ole suuria eroja, koska esimerkiksi tunnistamatta jäävien sananmuotojen osuudet ovat melko lähellä toisiaan. Morfo-ohjelma tosin tuotti syötetyistä sananmuodoista enemmän perusmuotoja, mutta se todennäköisimmin johtuu tässä tutkimuksessa käytetyistä Morfo-ohjelman asetuksista eli tarkoituksellisesta ylitulkitsevuudesta (luku 8.1).

Morfo-ohjelman tuottamat lisätulkinnat eivät kuitenkaan vaikuttaneet merkittävästi hakemistoon tallennettavien merkkijonojen määrään, vaan tämä oli jopa pienempi kuin Twol-ohjelman tuloksena saatava hakemiston merkkijonojen määrä (taulukko 11). Asia selittynee sillä, että Twol-ohjelmalta jäi tunnistamatta hiukan suurempi osuus syötetyistä sananmuodoista. Kun nämä tallennettiin sellaisinaan tunnistamattomien sananmuotojen hakemistoon, tämän hakemiston merkkijonojen määrä oli suurempi kuin Morfo-ohjelman jäljiltä. Näin myös muistitilaa tarvittaisiin enemmän kuin Morfo-ohjelmalla tuotetuissa hakemistoissa.

Yhden päivän tekstiaineisto sekä Twol- ja Morfo-ohjelmien väliset erot ovat kuitenkin niin pieniä, että eroavuudet voivat johtua mittausvirheistäkin. Perusteellisempi vertailu edellyttäisi suurempaa aineistoa sekä testihakujen tekemistä, jolloin voitaisiin tutkia, vaikuttavatko tällaiset hakemistojen erot myös hakutuloksiin.

Taulukko 11. Merkkijonojen ja osoitteiden määrä hakemistossa sekä hakemiston viemä muistitila lohkoina ja megatavuina.

	Merkki-jonojen määrä, kpl	Osoitteiden määrä, kpl	Muistitila lohkoina	Muistitila mega-tavuina
Twol	11 280	30 252	2 533	1,30
Morfo	10 828	30 221	2 445	1,25

8.6 Hakemisto käyttäjän silmin

Muistitilan säästeliäs käyttö oli aikoinaan tärkeää (Bell & Jones 1979), mutta asian merkitys on nyttemmin vähentynyt, kun tallennusmuisti on yhä halvempaa. Hakija ei sitä paitsi käytännössä välttämättä edes huomaa erilaisten hakemistoratkaisujen tehokkuus- ja tilankäyttöeroja. Sen sijaan hakuja tehtäessä ja erityisesti hakemistoja selattaessa eri hakemistotyyppien väliset erot ovat myös käyttäjän havaittavissa.

Hakemiston selaamista varten BASIS-hakujärjestelmässä on käytettävissä KATSO-komento, jonka parametriksi käyttäjä antaa haluamansa sanan (merkkijonon). Näin hakija voi halutessaan esimerkiksi tarkistaa, miten monta kertaa kukin sana (sananmuoto) on esiintynut hakemistossa ja muokata kyselyään tämän tiedon perusteella. Seuraavassa on käytetty esimerk-

Taulukko 12. Otos taivutusmuotohakemistosta (H1).

/ katso teksti=paperi**	/ katso teksti=peruskoulu**
JUTUT. termit	JUTUT. termit
A 64 TEKSTI=PAPERI	A 17 TEKSTI=PERUSKOULU
B 62 TEKSTI=PAPERIA	B 4 TEKSTI=PERUSKOULUA
C 1 TEKSTI=PAPERIAAN	C 5 TEKSTI=PERUSKOULUIHIN
D 1 TEKSTI=PAPERIAIKAKAUDELTA	D 2 TEKSTI=PERUSKOULUIKÄISTEN
E 1 TEKSTI=PAPERIALA	E 2 TEKSTI=PERUSKOULUISSA
F 1 TEKSTI=PAPERIALAA	F 2 TEKSTI=PERUSKOULUISTA
G 1 TEKSTI=PAPERIALALLA	G 7 TEKSTI=PERUSKOULUJEN
H 2 TEKSTI=PAPERIALAN	H 1 TEKSTI=PERUSKOULUKIRJALLE
I 1 TEKSTI=PAPERIANOMUKSIA	I 1 TEKSTI=PERUSKOULUKIRJOIHIN
J 1 TEKSTI=PAPERIARKISTA	J 1 TEKSTI=PERUSKOULUKIRJOILLA
K 1 TEKSTI=PAPERIARKKI	K 1 TEKSTI=PERUSKOULUKIRJOJEN
L 1 TEKSTI=PAPERIASIAT	L 1 TEKSTI=PERUSKOULULAIN
M 7 TEKSTI=PAPERIEN	M 2 TEKSTI=PERUSKOULULAINEN
N 1 TEKSTI=PAPERIEROTTELUN	N 2 TEKSTI=PERUSKOULULAISEN
O 1 TEKSTI=PAPERIEROTUS	O 1 TEKSTI=PERUSKOULULAISESTA
P 1 TEKSTI=PAPERIHAALAREIHIN	P 1 TEKSTI=PERUSKOULULAISET
Q 1 TEKSTI=PAPERIHIRVIÖ	Q 1 TEKSTI=PERUSKOULULAISETKIN
R 1 TEKSTI=PAPERIHIRVIÖN	R 1 TEKSTI=PERUSKOULULAISILLE
S 1 TEKSTI=PAPERIHUUH	S 1 TEKSTI=PERUSKOULULAISISTA
T 15 TEKSTI=PAPERIIN	T 1 TEKSTI=PERUSKOULULAISSA
U 1 TEKSTI=PAPERIJOHTAJA	U 1 TEKSTI=PERUSKOULULAISTA
V 1 TEKSTI=PAPERIJÄRJESTÖILLE	V 1 TEKSTI=PERUSKOULULAISTEN
W 2 TEKSTI=PAPERIJÄTTEEN	W 1 TEKSTI=PERUSKOULULINJOILLE
X 1 TEKSTI=PAPERIJÄTTEESTÄ	X 1 TEKSTI=PERUSKOULULLA
Y 1 TEKSTI=PAPERIKASA	Y 70 TEKSTI=PERUSKOULUN
Z 1 TEKSTI=PAPERIKASAN	Z 1 TEKSTI=PERUSKOULUNOPETTAJA

keinä paperi- ja peruskoulu-sanoja, jotka on katkaistu **-merkeillä. Tällöin BASIS näyttää hakemistosta aina kaikki hakuavainta järjestyksessä seuraavat 25 hakemiston merkkijonoa, riippumatta siitä, millä merkkijonolla nämä alkavat (eli ne eivät välttämättä täsmää hakusanan kanssa, sattuvat vain olemaan seuraavina aakkosjärjestyksessä).

Taivutusmuotohakemistossa (H1) käyttäjä tyypillisesti saa näkyvilleen hakusanan eri taivutusmuotoja sekä hakusanalla alkavia yhdyssanoja (taulukko 12). Usein yhdyssanat ja taivutusmuodot ovat hakemistossa lomittain vuorotellen niin, että käyttäjä ei pysty kovin näppärästi poimimaan vain tietyntyyppisiä hakemistosanoja (esimerkiksi pelkästään haetun sanan taivutusmuodot muttei yhdyssanoja). Sitä paitsi sanan kaikki taivutusmuodot eivät useinkaan mahdu yhtaikaa näyttöruudulle. Esimerkiksi paperi-sanan taivutusmuotoja esiintyy taulukon 12 vasemman sarakkeen kohdissa A-C, M ja T. Osa taivutusmuodoista jää edelleen näkymättömiin, esimerkiksi paperissa. Sitä paitsi osa taivutusmuodoista on hakemistossa jo ennen paperi-perusmuotoa, esimerkiksi papereita.

Taulukon 12 oikeanpuoleisesta sarakkeesta nähdään myös, että taivutusmuotohakemistossa peruskoulu-sanan genetiivimuotoja peruskoulun on enemmän (70 kappaletta) kuin nominatiivimuotoja peruskoulu (17 kappaletta). Tämä ei ole mitenkään poikkeuksellista, vaan nominatiivien määrä oli hakemistossa usein pienempi kuin genetiivien. On siis mahdollista, että dokumentissa esiintyy vain hakusanan genetiivimuoto, muttei lainkaan nominatiivia (eli perusmuotoa). Paljon käyttökelpoisia dokumentteja jää löytymättä, jos käyttäjä ei taivutusmuotohakemistosta hakiessaan muista katkaista hakusanaa.

Kun sananmuodot palautetaan perusmuotoon ennen kuin ne tallennetaan hakemistoon (H2), eri taivutusmuodot yhdistyvät yhteen perusmuotoon. Tällöin sanaluettelo on tiiviimpi kuin taivutusmuotohakemistossa eikä näyttö täyty yhden ja saman sanan eri muodoista (taulukko 13).

Perusmuotohakemistossa kuitenkin näkyy sanojen monitulkintaisuuden ongelma: homografisista ilmauksista voi syntyä huvittaviakin tulkintoja, kuten paperiarkista paperiarkki-perusmuodon lisäksi myös paperiarkinen. (Seikka)peräisempi-muodosta taas syntyy sekä adjektiivi -peräinen että verbi -peräis-empiä. Tällaiset hakemistosanat ovat omiaan hämmentämään käyttäjää ja voivat saada hänet epäilemään koko hakujärjestelmän kelvollisuut-

ta. Jotta epäkelpoja tulkintoja joutuisi hakemistoon mahdollisimman vähän, hakujärjestelmässä tulisi käyttää perusmuotoon palauttamisen jälkeen disambiguointiohjelmia ylimääräisten tulkintojen karsimiseksi.

Kun perusmuotoon palauttamisen lisäksi on jaettu yhdyssanat osiinsa ja osista tuotettu niiden yhdistelmät (H3), merkkijonojen määrä hakemistossa lisääntyy. Vaikka hakemistoa selattaessa näkyvät sekä perusmuodot että yhdyssanojen osat niiden joukossa (taulukko 14), näytölle mahtuu silti enemmän eri sanoja kuin taivutusmuotohakemistossa (H1).

Koska yhdyssanojen osat tallennettiin H3-hakemistossa niin, että osiin lisättiin tavuviiva osoittamaan niiden sijaintia yhdyssanassa, yhdyssanan keski- ja loppuosat tallentuivat eri kohtaan hakemistoa kuin alkuosat ja perussanat (taulukko 15). Käyttäjän on selattava näitä osia erikseen, mikä tietysti on

Taulukko 13. Otos perusmuotohakemistosta (H2).

/ katso teksti=paperi**	/ katso teksti=peruskoulu**
JUTUT. termit	JUTUT. termit
A 422 TEKSTI=PAPERI	A 106 TEKSTI=PERUSKOULU
B 1 TEKSTI=PAPERIAIKAKAUSI	B 2 TEKSTI=PERUSKOULUIKÄINEN
C 5 TEKSTI=PAPERIALA	C 1 TEKSTI=PERUSKOULUKIRJA
D 1 TEKSTI=PAPERIANOMUS	D 1 TEKSTI=PERUSKOULUKIRJO
E 1 TEKSTI=PAPERIARKINEN	E 9 TEKSTI=PERUSKOULULAINEN
F 2 TEKSTI=PAPERIARKKI	F 3 TEKSTI=PERUSKOULULAKI
G 1 TEKSTI=PAPERIASIA	G 1 TEKSTI=PERUSKOULULINJA
H 1 TEKSTI=PAPERIEROTTELU	H 2 TEKSTI=PERUSKOULUNOPETTAJA
I 1 TEKSTI=PAPERIEROTUS	I 1 TEKSTI=PERUSKOULUOPETUS
J 1 TEKSTI=PAPERIHAALARI	J 1 TEKSTI=PERUSKOULUPOHJA
K 1 TEKSTI=PAPERIHIRVIÖ	K 6 TEKSTI=PERUSKOULUPOHJAINEN
L 1 TEKSTI=PAPERIJOHTAJA	L 1 TEKSTI=PERUSKOULUTTAA
M 1 TEKSTI=PAPERIJÄRJESTÖ	M 15 TEKSTI=PERUSKOULUTUS
N 3 TEKSTI=PAPERIJÄTE	N 1 TEKSTI=PERUSKOULUTUSJAKSO
O 4 TEKSTI=PAPERIKASA	O 1 TEKSTI=PERUSKOULUTUSKAUSI
P 1 TEKSTI=PAPERIKASSI	P 1 TEKSTI=PERUSKOULUTUSTASO
Q 2 TEKSTI=PAPERIKAUPPA	Q 1 TEKSTI=PERUSKOULUVÄKI
R 1 TEKSTI=PAPERIKAUPPIAS	R 1 TEKSTI=PERUSKULUTUS
S 1 TEKSTI=PAPERIKEMIA	S 3 TEKSTI=PERUSKUNNOSTUS
T 1 TEKSTI=PAPERIKEMIKAALITEHDAS	T 4 TEKSTI=PERUSKUNTA
U 1 TEKSTI=PAPERIKOKO	U 7 TEKSTI=PERUSKUNTO
V 38 TEKSTI=PAPERIKONE	V 1 TEKSTI=PERUSKUNTOHARJOITTELU
W 1 TEKSTI=PAPERIKONEENTEKIJÄ	W 4 TEKSTI=PERUSKUNTOKAUSI
X 1 TEKSTI=PAPERIKONEHANKE	X 14 TEKSTI=PERUSKURSSI
Y 1 TEKSTI=PAPERIKONEINEN	Y 1 TEKSTI=PERUSKURSSITUS
Z 1 TEKSTI=PAPERIKONEINVESTOINTI	Z 1 TEKSTI=PERUSKUUNNELMA

hankalaa. Tosin on muistettava, että näitä ilmauksia ei taivutusmuotohakemistossa ei ole lainkaan eikä niitä siten voi siinä edes selata. Mikäli hakujärjestelmä sallii hakusanan katkaisun vasemmalta, yhdyssanan loppuosia voidaan taivutusmuotohakujärjestelmässäkin hakea, mutta se vaatii paljon tietokoneresursseja.

Ne sananmuodot, joita Morfo ei tunnistanut, tallennettiin tunnistamattomien sananmuotojen hakemistoon. Myös tätä hakemistoa voitiin selata BASIS-järjestelmän KATSO-komennon avulla. Paperi ja peruskoulu eivät tietenkään löydy sellaisinaan tästä hakemistosta, koska Morfo tunnistaa ne, mutta yhdenmukaisuuden vuoksi nämä samat sanat annettiin myös tässä tapauksessa selauksen lähtökohdaksi (taulukko 16).

Taulukko 14. Otos ositetusta perusmuotohakemistosta (H3).

/ katso teksti=paperi**	/ katso teksti=peruskoulu**
JUTUT. termit	JUTUT. termit
A 410 TEKSTI=PAPERI	A 106 TEKSTI=PERUSKOULU
B 350 TEKSTI=PAPERI-	B 26 TEKSTI=PERUSKOULU-
C 1 TEKSTI=PAPERIAIKA-	C 2 TEKSTI=PERUSKOULUIKÄINEN
D 1 TEKSTI=PAPERIAIKAKAUSI	D 1 TEKSTI=PERUSKOULUKIRJA
E 5 TEKSTI=PAPERIALA	E 1 TEKSTI=PERUSKOULUKIRJO
F 1 TEKSTI=PAPERIANOMUS	F 9 TEKSTI=PERUSKOULULAINEN
G 1 TEKSTI=PAPERIARKINEN	G 3 TEKSTI=PERUSKOULULAKI
H 2 TEKSTI=PAPERIARKKI	H 1 TEKSTI=PERUSKOULULINJA
I 1 TEKSTI=PAPERIASIA	I 2 TEKSTI=PERUSKOULUNOPETTAJA
J 1 TEKSTI=PAPERIEROTTELU	J 1 TEKSTI=PERUSKOULUOPETUS
K 1 TEKSTI=PAPERIEROTUS	K 1 TEKSTI=PERUSKOULUPOHJA
L 1 TEKSTI=PAPERIHAALARI	L 6 TEKSTI=PERUSKOULUPOHJAINEN
M 1 TEKSTI=PAPERIHIRVIÖ	M 1 TEKSTI=PERUSKOULUTTAA
N 1 TEKSTI=PAPERIJOHTAJA	N 15 TEKSTI=PERUSKOULUTUS
O 1 TEKSTI=PAPERIJÄRJESTÖ	O 3 TEKSTI=PERUSKOULUTUS-
P 3 TEKSTI=PAPERIJÄTE	P 1 TEKSTI=PERUSKOULUTUSJAKSO
Q 2 TEKSTI=PAPERIKASA	Q 1 TEKSTI=PERUSKOULUTUSKAUSI
R 1 TEKSTI=PAPERIKASSI	R 1 TEKSTI=PERUSKOULUTUSTASO
S 2 TEKSTI=PAPERIKAUPPA	S 1 TEKSTI=PERUSKOULUVÄKI
T 1 TEKSTI=PAPERIKAUPPIAS	T 1 TEKSTI=PERUSKULUTUS
U 1 TEKSTI=PAPERIKEMI-	U 3 TEKSTI=PERUSKUNNOSTUS
V 1 TEKSTI=PAPERIKEMIA	V 4 TEKSTI=PERUSKUNTA
W 1 TEKSTI=PAPERIKEMIKAALI-	W 7 TEKSTI=PERUSKUNTO
X 1 TEKSTI=PAPERIKEMIKAALITEHDAS	X 5 TEKSTI=PERUSKUNTO-
Y 1 TEKSTI=PAPERIKOKO	Y 1 TEKSTI=PERUSKUNTOHARJOITTELU
Z 38 TEKSTI=PAPERIKONE	Z 4 TEKSTI=PERUSKUNTOKAUSI

Tunnistamattomat sananmuodot voidaan karkeasti jakaa seuraaviin ryhmiin:

- kirjoitusvirheet (peruskymyksistä, perustamaaan)
- sanakirjasta puuttuvat erisnimet; sekä vierasperäiset (Papoulkos) että suomenkieliset (Raatikko)
- sanakirjasta jostain syystä puuttuvat yleisnimet; sekä vierasperäiset että suomenkieliset (takki)

Näistä ongelmatapauksista ainoastaan suomenkieliset yleisnimet ovat ryhmä, jonka periaatteessa pitäisi sisältyä sanakirjaan kokonaan. Sen sijaan muita ongelmatapauksia ei voida lisätä sanakirjaan - kirjoitusvirheet eivät sinne kuulu ja uusia nimiä taas syntyy jatkuvasti niin paljon, että sanakirjaa on mahdoton pitää ajan tasalla.

Hakujärjestelmät pitää siis rakentaa sillä periaatteella, että teksteissä esiin-

Taulukko 15. Otos ositetusta perusmuotohakemistosta (H3).

/ katso teksti=-paperi**	/ katso teksti=-peruskoulu**
JUTUT. termit	JUTUT. termit
A 154 TEKSTI=-PAPERI	A 4 TEKSTI=-PERUSTA
B 26 TEKSTI=-PAPERI-	B 13 TEKSTI=-PERUSTAA
C 2 TEKSTI=-PAPERIERÄ	C 3 TEKSTI=-PERUSTAINEN
D 1 TEKSTI=-PAPERIKAPASITEETTI	D 1 TEKSTI=-PERUSTAMINEN
E 1 TEKSTI=-PAPERIKASSI	E 84 TEKSTI=-PERUSTE
F 2 TEKSTI=-PAPERIKONE	F 3 TEKSTI=-PERUSTE-
G 1 TEKSTI=-PAPERIKONE-	G 29 TEKSTI=-PERUSTEINEN
H 1 TEKSTI=-PAPERIKONETILAUS	H 1 TEKSTI=-PERUSTEISESTI
I 1 TEKSTI=-PAPERIKUORI	I 12 TEKSTI=-PERUSTELU
J 1 TEKSTI=-PAPERILAJI	J 2 TEKSTI=-PERUSTELUU
K 1 TEKSTI=-PAPERIMARKKINAT	K 1 TEKSTI=-PERUSTETOIMI-
L 10 TEKSTI=-PAPERINEN	L 1 TEKSTI=-PERUSTETOIMIKUNTA
M 1 TEKSTI=-PAPERINÄYTE	M 19 TEKSTI=-PERUSTIE
N 3 TEKSTI=-PAPERIRULLA	N 1 TEKSTI=-PERUUKKI
O 2 TEKSTI=-PAPERIRYHMÄ	O 1 TEKSTI=-PERUUTUS
P 7 TEKSTI=-PAPERITEHDAS	P 140 TEKSTI=-PERÄ
Q 3 TEKSTI=-PAPERITEOLLIS-	Q 32 TEKSTI=-PERÄ-
R 3 TEKSTI=-PAPERITEOLLISUUS	R 81 TEKSTI=-PERÄINEN
S 3 TEKSTI=-PAPERITEOLLISUUSI	S 2 TEKSTI=-PERÄIS-
T 1 TEKSTI=-PAPERITON	T 1 TEKSTI=-PERÄISEMPIÄ
U 1 TEKSTI=-PAPERITUOTANTO	U 3 TEKSTI=-PERÄISESTI
V 3 TEKSTI=-PAPERIYHTIÖ	V 1 TEKSTI=-PERÄISIMPI
W 2 TEKSTI=-PAPISTO	W 11 TEKSTI=-PERÄISYYS
X 4 TEKSTI=-PAPPA	X 1 TEKSTI=-PERÄKÄRRY
Y 7 TEKSTI=-PAPPEUS	Y 1 TEKSTI=-PERÄLAATIKKO
Z 1 TEKSTI=-PAPPEUS-	Z 1 TEKSTI=-PERÄMIES

tyy ilmauksia, joita ei voi palauttaa perusmuotoon. Tällaisten ilmausten hakeminen on toteutettava eri tavalla kuin sellaisten sanojen, jotka voidaan ongelmitta perusmuotoistaa (ks. luku 10).

Kun hakemisto sisältää sanojen perusmuodot, hakemistosanojen selaaminen on periaatteessa yksinkertaisempaa kuin taivutusmuotohakemistossa. Yhdessä suhteessa taivutusmuotohakemisto kuitenkin on parempi: siinä kaikki selattavat sananmuodot ovat yhdessä hakemistossa. Käyttäjälle kahden eri hakemiston (tai useiden eri tunnuskoodien) käyttö tietysti on hankalaa. Jos hän esimerkiksi haluaisi hakemiston kautta selvittää, millä tavoin Peruzzi on teksteissä kirjoitettu, tätä tietoa ei löydy perusmuotohakemistosta, josta käyttäjä todennäköisesti sitä ensimmäiseksi etsii. Tämän jälkeen käyttäjän

Taulukko 16. Otos tunnistamattomien sananmuotojen hakemistosta. Tekstissä esiintyneet tunnistamattomat ilmaukset on merkitty koodilla ZT, jotta ne voitaisiin erottaa Morfo-ohjelman tunnistamista sanoista.

1/ katso zt=paperi**	/ katso zt=peruskoulu**
JUTUT. termit	JUTUT. termit
A 1 ZT=PAPERIHUUH	A 1 ZT=PERUSKYMYKSISTÄ
B 1 ZT=PAPERIKONETETAALLA	B 1 ZT=PERUSLAISOPISKELIJA
C 1 ZT=PAPERILIIT	C 1 ZT=PERUSLEIRITYKSESTÄ
D 1 ZT=PAPERILIITTON	D 1 ZT=PERUSLIVERPOOLILAINEN
E 1 ZT=PAPERILITOSSA	E 1 ZT=PERUSNAHKATAKKI
F 7 ZT=PAPERITEHT	F 1 ZT=PERUSOIKEUSÄÄNNÖSTÖN
G 1 ZT=PAPERNI	G 1 ZT=PERUSPEKTIIVEJÄ
H 3 ZT=PAPERNY	H 1 ZT=PERUSSAVOLAINEN
I 1 ZT=PAPEROIDES	I 1 ZT=PERUSTAMAAAN
J 1 ZT=PAPETERIES	J 1 ZT=PERUSTAMISUUNNITELMAAN
K 2 ZT=PAPI	K 1 ZT=PERUSTEEELLA
L 1 ZT=PAPIERFABRIEKIN	L 1 ZT=PERUSTEELISESTI
M 1 ZT=PAPILA	M 1 ZT=PERUSTEELLLISESTI
N 1 ZT=PAPILLON	N 1 ZT=PERUSTEL
O 1 ZT=PAPILLONISTA	O 1 ZT=PERUSTEPSILÄINEN
P 1 ZT=PAPILLONS	P 1 ZT=PERUSTETIIN
Q 17 ZT=PAPINK	Q 1 ZT=PERUSTETTOMASTA
R 1 ZT=PAPLO	R 1 ZT=PERUSTETTUTYÖVÄENLIITTO
S 2 ZT=PAPO	S 1 ZT=PERUSTUKIMUS
T 1 ZT=PAPOULKOS	T 1 ZT=PERUSTUSLAKIAVALIOKUNNALLA
U 1 ZT=PAPP	U 1 ZT=PERUSVINK
V 1 ZT=PAPPAS	V 1 ZT=PERUTTAMAAN
W 1 ZT=PAPPEJ	W 1 ZT=PERUUTETIIN
X 1 ZT=PAPRUT	X 1 ZT=PERUUTTAAMINEN
Y 1 ZT=PAPUKA	Y 1 ZT=PERUUTTOMATTOMISTA
Z 1 ZT=PAPULAN	Z 1 ZT=PERUZZI

täytyy siirtyä tunnistamattomien sanojen hakemistoon (tai varustaa hakusana tunnistamattomien sanojen koodilla) ja tehdä sama selaus uudelleen tästä hakemistosta.

Tämä merkitsee sitä, että (ositetun) perusmuotohakemiston käyttöliittymä on kehitettävä sellaiseksi, että eri hakemistot käyttäjän näkökulmasta näyttävät yhtenäiseltä. Esimerkiksi ositetussa perusmuotohakemistossa yhdyssanojen eri osat (kuten osa, -osa, -osa-, osa-) pitäisi haluttaessa saada aakkostumaan samaan paikkaan.

9 PERUSMUOTOISET HAKUSANAT VAKIOKYSELYISSÄ

Tutkimuksen kyselyt 1 - 30 olivat **vakiokyselyjä**, joissa suomen kielen morfologisten tulkintaohjelmien soveltaminen ei tuottanut erityisiä ongelmia. Alkuperäisistä 30 vakiokyselystä jouduttiin kuitenkin muista syistä karsimaan pois neljä: Kysely numero 18 jätettiin tutkimusjoukosta pois, koska tutkimusympäristössä T3 (seulonta) sitä ei teknisistä syistä saatu suoritettua. Kyselyissä 22 ja 23 taas tulosjoukkoihin ei missään tutkimusympäristössä millään kyselytyypillä löydetty yhtään relevanttia artikkelia. Kyselyssä 30 yhdyssanan oikea jakokohta oli ongelmallinen, joten varmuuden vuoksi sekin poistettiin tutkimusjoukosta. Osumien lukumäärä sekä hakutuloksen saantiarvo ja tarkkuusarvo laskettiin siis 26 vakiokyselyn **perusjoukosta** kaikissa tutkimusympäristöissä.

Lisäksi erotettiin kaksi suppeampaa **osajoukkoa**, joista toisessa tutkittiin erityisesti johdoksien ja toisessa erityisesti yhdyssanojen osien käyttäytymistä.

Tapauksia, joissa hakusanan rinnalle kyselyyn voitiin lisätä aitoja **johdosperheen** jäseniä, oli 8 kappaletta - hakupyynnöissä 6, 7, 8, 9, 10, 11, 13 ja 20. Tämä osajoukko nimettiin **johdososajoukoksi**. Muissa hakupyynnöissä annettiin johdoskyselylle (AB), yhdistelmäkyselylle (ABC), osien johdoskyselylle (ABab) ja osien yhdistelmäkyselylle (ABCabc) suoraan niitä edeltävän tason (eli A-, AC-, Aa- tai ACac-kyselyn) saanti- ja tarkkuusarvo. Tämän katsottiin vastaavan tilannetta, jossa kyselyjä laajennetaan poimimalla hakutesauruksesta johdokset kaikille hakusanoille, joille sellaisia löytyy - jos johdoksia ei ole tai niillä ei saada yhtään uutta osumaa, tilanne on sama kuin ennen laajennusta.

Liitteiden 7 - 12 taulukoissa johdososajoukon kyselyt on merkitty tunnuksella J. Osumien lukumäärä sekä saanti- ja tarkkuusarvot on merkitty siten, että kun kyselyä voitiin aidosti laajentaa johdosperheellä, arvot merkittiin lihavoituna. Ne kyselyt, joissa johdoksia ei ollut ja joissa lukemat siis siirrettiin edeltävältä tasolta, on merkitty taulukkoihin normaalilla, lihavoimattomalla tekstillä.

Mikäli taas hakusana oli **yhdyssana, joka voitiin jakaa osiinsa**, kyselyä laajennettiin näillä osilla ja niiden johdosperheen jäsenillä. Tällaisia tapauk-

sia oli kaikkiaan 14 kappaletta eli hakupyynnöt 3, 4, 5, 11, 12, 13, 14, 15, 16, 17, 19, 20, 26 ja 29. Näistä viidessä kävi niin, että kyselyjen tuottamien tulosjoukkojen koko oli niin suuri, että tulosjoukoista oli tehtävä otanta. Luvussa 7.9 selostettujen epävarmuustekijöiden vuoksi katsottiin kuitenkin varmemmaksi jättää nämä otantakyselyt pois vertailulaskelmista. Näin ollen saanti- ja tarkkuuslaskelmat sekä merkitsevyystestit tehtiin vain niille kyselyille, joista otantaa ei tehty. Tällaiset kyselyt oli laadittu yhdeksästä hakupyynnöstä: 4, 5, 11, 15, 16, 17, 19, 20 ja 26. Nämä yhdessä muodostivat **yhdyssanaosajoukon**. Yhdyssanaosajoukon kyselyt on liitteissä 7 - 12 merkitty tunnuksella Y.

Kun jokaisen kyselyn tulosjoukosta oli laskettu dokumenttien lukumäärä sekä saanti- ja tarkkuusarvot, laskettiin näistä kyselykohtaisista tuloksista kunkin kyselytyypin keskimääräinen dokumenttien lukumäärä sekä saanti- ja tarkkuusarvot. Näin saatuja keskiarvoja ei kuitenkaan tulisi pitää kyselytyypin absoluuttisena tehokkuusarvona, vaan sitä on tulkittava vain **suhteessa tämän tutkimuksen** muiden kyselytyyppien ja tutkimusympäristöjen vastaaviin arvoihin.

Yksittäisen kyselytyypin saanti- ja tarkkuusarvojen suora vertailu toisten tutkimusten saanti- ja tarkkuusarvoihin (ks. esimerkiksi Sormunen 1994, s. 137 - 138) ei ole järkevää siksi, että FULLTEXT-projektin kyselyt sellaisinaan eivät olisi optimaalisia kyselyjä todellisissa tiedonhakutilanteissa. Niistä puuttuvat hakusanojen synonyymit, rinnakkaiset ilmaukset sekä ylä- ja alakäsitteet, joilla normaalisti kasvatetaan haun saantia.

Myöskään operaattorien vaikutusta kyselyjen tuloksiin ei tässä tutkimuksessa syvällisesti analysoitu ja pyritty optimoimaan. Esimerkiksi Tenopir ja Ro (1990) esittävät, että tekstihaussa kappaleoperaattorin käyttö tuottaa sekä saannin että tarkkuuden kannalta optimaalisen tuloksen. Kristensenin (1993) mukaan taas kappaleoperaattori vain huonontaa saantia verrattuna JA-operaattorin käyttöön. Toisaalta Sormusen (2000) tutkimuksessa JA-operaattorit tuottivat yhtä hyviä tuloksia kuin läheisyysoperaattorit, kun kyselyjen tarkkuustaso oli jo valmiiksi korkea. FULLTEXT-projektissa operaattoreiden hienosäätö ja esimerkiksi kappaleoperaattorin käyttö ei ollut mahdollista, vaan hakusanat oli yhdistettävä joko JA- tai virkeoperaattorilla.

Tulosten tulkinnassa on siis otettava huomioon, että FULLTEXT-projektin varsinainen tavoite ei ollut maksimaalisesti hyödyntää BASIS-hakujärjestelmän ominaisuuksia, vaan toteuttaa käytännössä sellaisia hakemisto- ja kyselyvaihtoehtoja, joissa morfologisten tulkintaohjelmien tuotoksia ja niiden vaikutuksia hakutuloksiin voitaisiin parhaiten vertailla ja analysoida.

Seuraavissa alaluvuissa kuvataan yksityiskohtaisemmin eri muuttujien vaikutuksia eri tutkimusympäristöissä. Aluksi kutakin tutkimusympäristöä T2 - T5, joissa sovellettiin suomen kielen morfologisia tulkintaohjelmia, vertaillaan T1-tutkimusympäristön kanssa. Lopuksi viimeisessä alaluvussa 9.6 tarkastellaan kaikkia tutkimusympäristöjä yhdessä.

9.1 Taivutusmuotohakemistoon perustuva tutkimusympäristö

T1-tutkimusympäristössä hakija eli tutkija katkaisi hakusanat siten, että niillä löydettiin hakemistosta itse hakusanojen lisäksi myös hakusanojen koko johdosperhe (esimerkiksi autoverotus -> hakusana *autovero**). Jos hakusana oli yhdyssana, se voitiin jakaa osiinsa (esimerkiksi *auto** JA *vero**). Katkaistut hakusanat siis täsmäsivät hakemistossa varsinaisiin hakusanoihin, niiden johdoksiin sekä hakusanalla alkaviin yhdyssanoihin.

Yleisvaikutelmana on, että T1-ympäristössä hakutulosten **saantiarvo** nousi korkeaksi verrattuna muiden tutkimusympäristöjen kyselyjen hakutuloksiin, mutta samalla **tarkkuus** kuitenkin kärsi; erityisesti näin kävi osien yhdistelmäkyselyssä ABCabc (liite 14). Osien yhdistelmäkyselyn huono tarkkuus selittyy siitä, että yhdyssanojen osana olevien perussanojen katkaisu tuottaa hyvin lyhyitä hakusanoja. Lyhyitä hakusanoja käytettäessä haun tuloksena saadaan paljon osumia, mutta epärelevanttien dokumenttien osuus on suuri: monissa dokumenteissa on vain sattumalta sananmuoto, joka alkaa samalla merkkijonolla kuin hakusana.

9.2 Hakijan katkaisemien ja automaattisesti katkaistujen hakusanojen vertailu

Aluksi kuvataan T2-tutkimusympäristön tulokset, kun hakusanojen automaattiseen katkaisuun käytettiin Finstems-ohjelmaa. Sen jälkeen tarkastellaan, miten Hahmotin-ohjelman tuottamat vartalot poikkesivat Finstems-ohjelman vartaloista ja miten erot vaikuttivat hakutuloksiin.

9.2.1 Finstemsin tuottamilla vartaloilla hakeminen

9.2.1.1 Perusjoukko

Yhdyssanakysely

Kun hakusanat katkaistiin Finstemsin avulla (taulukko 17, kyselytyyppi AC), tulosjoukon keskimääräinen **koko** jäi pienemmäksi kuin perinteisellä yhdistelmäkyselyllä (ABC) haettaessa.

Samalla hakutuloksen **saantikin** jäi alemmaksi: JA-operaattoria käytettäessä yhdyssanakyselyn (AC) suhteellinen saanti oli 8 prosenttiyksikköä ja virkeoperaattoria käytettäessä vajaat 6 prosenttiyksikköä alempi kuin T1-ympäristön yhdistelmäkyselyn.

Yhdyssanakyselyn (AC) ja perinteisen yhdistelmäkyselyn (ABC) saantiarvojen välinen ero oli Friedmanin merkitsevyystestin perusteella sekä JA- että virkeoperaattoria käytettäessä tilastollisesti merkitsevä merkitsevyystasolla 0.01. Sparck Jonesin käytännön merkitsevyys -mittarilla mitattuna hakutulosten välinen ero oli huomattava, siis yli viiden ja alle kymmenen prosenttiyksikön verran (liite 14, luku 1).

Aivan vastaava tulos saadaan verrattaessa yhdyssanakyselyä saman T2-ympäristön yhdistelmäkyselyn (ABC) kanssa: sekä JA- että virkeoperaattorilla saantiarvojen väliset erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.01 ja käytännössä huomattavat.

Taulukko 17. T1- ja T2-ympäristöissä (Finstems) kyselytyypeillä AC ja ABC saatujen tulosjoukkojen väliset erot perusjoukossa (N =26).

Operaattori	Kyselytyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T2**	**T1**	**T2**	**T1**	**T2**
JA	AC		12,5		63,9		71,1
	ABC	**20,0**	**19,8**	**72,0**	**71,8**	**68,0**	**68,7**
Virke	AC		8,3		54,2		77,3
	ABC	**11,3**	**11,2**	**60,1**	**60,0**	**76,6**	**76,6**

AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat

Hypoteesin 2 väite saannin huononemisesta siis pitää paikkansa, koska saantiarvojen välinen ero on paitsi tilastollisesti merkitsevä, myös käytännössä huomattava. Hakusanan johdokset siis ovat tärkeitä haun saannin kannalta: jos dokumentissa ei ole itse hakusanaa, vaan pelkästään hakusanan johdosperheen jokin tai joitakin muita jäseniä, dokumentti jää pois yhdyssanakyselyn tulosjoukosta, mistä seuraa huonompi saanti.

Automaattisesti katkaistuja hakusanoja käytettäessä yhdyssanakyselyn (AC) **tarkkuus** oli parempi kuin hakijan itse katkaisemia hakusanoja käytettäessä (ABC), mutta ero näiden kyselytyyppien välillä oli JA-operaattoria käytettäessä 3 prosenttiyksikköä ja virkeoperaattorilla vain vajaan prosenttiyksikön.

Tarkkuusarvojen väliset erot olivat JA-operaattoria käytettäessä tilastollisesti merkitseviä merkitsevyystasolla 0.025. Virkeoperaattoria käytettäessä tarkkuusarvojen väliset erot eivät olleet tilastollisesti merkitseviä. Myöskään Sparck Jonesin mittarilla mitaten tarkkuusarvojen välisillä eroilla ei ollut käytännössä merkitystä.

Hypoteesin 2 väite tarkkuuden paranemisesta saa vain lievästi tukea, eivätkä tarkkuusarvojen erot ole yhtä selviä kuin saantiarvoja vertailtaessa.

Perusjoukossa automaattinen katkaisu siis alensi tulosjoukkojen saantiarvoja huomattavasti verrattuna perinteisellä tavalla toteutettuun kyselyyn, mutta tarkkuusarvot nousivat vain hiukan. Saantiarvojen alenemisen tilastollinen todennäköisyys oli selvempi kuin tarkkuusarvojen nousemisen todennäköisyys.

Yhdistelmäkysely

Kun T2-ympäristön yhdyssanakyselyä laajennettiin johdosperheellä eli muodostettiin yhdistelmäkysely (ABC), T1- ja T2-ympäristöjen väliset erot tasoittuivat. Tulosjoukkojen koot olivat näissä kahdessa ympäristössä lähes samat, samaten saanti- ja tarkkuusarvot. T1- ja T2-ympäristöjen saanti- ja tarkkuusarvojen väliset erot eivät olleet tilastollisesti merkitseviä.

Hypoteesin 3 väite saannin samuudesta siis piti paikkansa, mutta väite paremmasta tarkkuudesta ei T2-ympäristön perusjoukossa saanut tukea.

Automaattinen katkaisu on perinteiseen hakutapaan eli hakijan itse katkaisemiin hakusanoihin verrattuna vaivattomampi, kun hakijan ei itse tarvitse

päätellä oikeaa katkaisukohtaa. Yhtä hyvä saanti kuin perinteisellä hakutavalla saavutetaan kuitenkin vain, kun johdokset otetaan huomioon.

9.2.1.2 Johdososajoukko

Seuraavaksi vertailut tehtiin johdososajoukossa eli siinä kahdeksan kyselyn osajoukossa, jossa hakusanoille löytyi johdoksia ja peruskyselyn laajentaminen johdosperheellä tuotti tulosjoukkoon uusia dokumentteja (taulukko 18; liite 14, luku 2).

Yhdyssanakysely

Tässä osajoukossa T1- ja T2-tutkimusympäristöjen väliset erot tulivat näkyviin perusjoukkoa selvemmin. Ensimmäinen selvä ero oli, että yhdyssanakyselyn (AC) tuloksena saadut tulosjoukot olivat selvästi pienempiä kuin perinteisen yhdistelmäkyselyn (ABC) tuloksena saadut tulosjoukot.

Kun hakusanat oli katkaistu automaattisesti, yhdyssanakyselyn (AC) **saanti** oli JA-operaattoria käytettäessä 26 prosenttiyksikköä ja virkeoperaattoria käytettäessä 19 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn. Sekä JA- että virkeoperaattoria käytettäessä yhdyssanakyselyn ja perinteisen yhdistelmäkyselyn saantiarvojen välinen ero oli Friedmanin testin mukaan tilastollisesti merkitsevä merkitsevyystasolla 0.01. Saantiarvojen välinen ero oli käytännössä olennainen (> 10 prosenttiyksikköä). Hypoteesi 2 siis piti saannin suhteen täysin paikkansa.

T2-ympäristön yhdyssanakyselyn (AC) **tarkkuus** oli JA-operaattoria käy-

Taulukko 18. T1- ja T2-ympäristöissä (Finstems) kyselytyypeillä AC ja ABC saatujen hakutulosten väliset erot johdososajoukossa (N = 8).

Operaattori	Kyselytyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T2**	**T1**	**T2**	**T1**	**T2**
JA	AC		17,6		57,1		54,7
	ABC	**41,3**	**41,3**	**82,7**	**82,7**	**46,9**	**46,9**
Virke	AC		6,3		37,3		66,5
	ABC	**15,9**	**15,9**	**56,2**	**56,2**	**64,3**	**64,3**

AC Yhdyssanakysely
ABC Yhdistelmäkysely
T1 Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2 Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat

tettäessä noin 8 prosenttiyksikköä ja virkeoperaattoria käytettäessä vähän yli 2 prosenttiyksikköä parempi kuin vastaavan perinteisen kyselyn.

Tarkkuusarvojen väliset erot olivat JA-operaattoria käytettäessä tilastollisesti merkitseviä merkitsevyystasolla 0.025. Virkeoperaattoria käytettäessä tarkkuusarvojen väliset erot eivät olleet tilastollisesti merkitseviä. Sparck Jonesin mittarilla ero oli JA-operaattorin tapauksessa käytännössä huomattava.

Hypoteesin 2 väite automaattisen katkaisun tuottamista paremmista tarkkuusarvoista verrattuna hakijan katkaisemien hakusanojen tuottamien tulosjoukkojen tarkkuusarvoihin siis sai tukea, mutta varsin lievästi. Ero ei ollut yhtä selvä kuin saantiarvojen tapauksessa.

Yhdistelmäkysely

Kun T2-ympäristön yhdyssanakyselyjä laajennettiin tässä kahdeksan kyselyn osajoukossa lisäämällä johdosperheet kyselyihin (eli muodostamalla yhdistelmäkysely ABC), T1-ja T2-ympäristöjen väliset erot katosivat kokonaan. Tämä tulos johtuu siitä, että Finstems-ohjelma katkaisi verbivartalot melko lyhyiksi. Kun lyhyet verbivartalot lisättiin kyselyyn, niillä saatiin samat hakemistosanat kuin hakijan itse katkaisemilla lyhyillä hakusanoilla.

Muut johdosperheen jäsenet olivatkin käytännössä tarpeettomia kyselyssä, kun katkaistut verbivartalot muita lyhyempinä palauttivat tietenkin kaikki nekin muodot, jotka pitemmillä vartaloilla olisi saatu (esimerkiksi tutkia-verbistä saadulla *tutki**-hakusanalla löydetään kaikki tutkimus-hakemistosanan eri taivutusmuodot). Perinteisen yhdistelmäkyselyn mukaisiin tuloksiin päästäisiin siis helposti siten, että kullekin hakusanalle etsittäisiin tai muodostettaisiin sen verbijohdos tai verbikantasana, joka sitten katkaistaisiin Finstems-ohjelman avulla. Tosin tarkkuuskin sitten jäisi yhtä alhaiseksi kuin perinteisen yhdistelmäkyselyn.

Hypoteesin 3 väittämä saannin samanlaisuudesta siis piti paikkansa. Sen sijaan väite T2-ympäristön paremmista tarkkuusarvoista ei saanut tukea.

Johdososajoukossa siis näkyy perusjoukkoakin selvemmin, että hakusanojen automaattinen katkaisu ei sellaisenaan ole toimiva ratkaisu, jos johdokset unohtuvat tai jätetään kyselystä pois: hakutulosten tarkkuus ei kovinkaan paljon parane verrattuna perinteiseen yhdistelmähakuun, kun taas saanti huononee olennaisesti. Saantiarvojen huomattava aleneminen juuri

johdososajoukossa on sinänsä loogista, koska tämän osajoukon kyselyissä ovat mukana ne hakusanat, joilla on rinnakkaisia johdosilmauksia - olisi outoa, jos niiden jättäminen pois kyselyistä ei alentaisi saantiarvoja. Sen sijaan on mielenkiintoista, että vertailtujen kyselytyyppien tarkkuusarvojen välinen ero jääkin odotettua pienemmäksi.

9.2.1.3 Yhdyssanaosajoukko

Perinteistä yhdistelmähakua ja automaattista katkaisua vertailtiin myös yhdyssanaosajoukossa eli siinä yhdeksän kyselyn osajoukossa, jossa yhdyssanat voitiin jakaa osiinsa (taulukko 19; liite 14, luku 3).

Hakutuloksissa näkyy selvä ero, kun yhdyssanat ositetaan ja osat lisätään kyselyyn verrattuna kokonaisilla yhdyssanoilla tehtyihin kyselyihin. Tämä ei tietenkään ole yllättävää, koska tässä osajoukossa yhdyssanojen voidaan olettaa vaikuttavan hakutulokseen enemmän kuin johdososajoukossa, jossa taas olennainen muutos tapahtuu, kun kyselyä laajennetaan johdoksilla.

Kun esimerkiksi yhdistelmäkyselystä (ABC) siirryttiin osien yhdistelmäkyselyyn (ABCabc), haun **saanti** kasvoi sekä T1- että T2-ympäristöissä JA-operaattoria käytettäessä yli 32 prosenttiyksikköä ja virkeoperaattorillakin

Taulukko 19. T1- ja T2-ympäristöissä (Finstems) kyselytyypeillä ACac ja ABCabc saatujen hakutulosten väliset erot yhdyssanaosajoukossa (N = 9).

Operaattori	Kyselytyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T2**	**T1**	**T2**	**T1**	**T2**
JA	AC		7,0		54,0		78,4
	ABC	8,0	7,4	55,7	55,7	76,1	78,2
	ACac		21,9		83,1		47,1
	ABCabc	**25,8**	**25,0**	**88,1**	**88,1**	**42,1**	**43,0**
Virke	AC		5,8		47,1		76,7
	ABC	6,1	6,1	48,8	48,8	80,0	80,0
	ACac		8,1		61,0		78,1
	ABCabc	**8,6**	**8,4**	**62,8**	**62,8**	**81,1**	**81,4**

AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
ACac	Osien yhdyssanakysely
ABCabc	Osien yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstemsin katkaisemat hakusanat

14 prosenttiyksikköä. Tällöin näiden saantiarvojen väliset erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.01 ja käytännössä olennaisia.

Itse asiassa kaikkien niiden kyselytyyppien, joissa yhdyssanoja ei ollut ositettu (siis T1/ABC, T2/AC, T2/ABC), saanti erosi merkitsevyystasolla 0.01 kaikkien niiden kyselytyyppien saannista, joissa yhdyssanat oli ositettu (eli T1/ABCabc, T2/ACac, T2/ABCabc). Tämä päti sekä JA- että virkeoperaattoria käytettäessä. Myös Sparck Jonesin mittarilla mitaten kaikkien ensinmainittujen kyselytyyppien ja yhdyssanojen osia sisältävien kyselytyyppien saantiarvojen väliset erot olivat käytännössä olennaiset.

Osien yhdyssanakyselyssä (ACac) automaattisen katkaisun saanti jäi JA-operaattoria käytettäessä 5 prosenttiyksikköä ja virkeoperaattoria käytettäessä vajaat 2 prosenttiyksikköä alemmaksi kuin perinteisen osien yhdistelmäkyselyn (T1/ABCabc). Virkeoperaattorin tapauksessa ero ei ollut tilastollisesti merkitsevä, JA-operaattoria käytettäessäkään ei varsinaisesti (ero oli merkitsevä vain tasolla 0.1). - Johdoksilla ei siis yhdyssanaosajoukossa ole saannin kannalta niin suurta merkitystä kuin perus- ja johdososajoukossa.

Kun kyselyä laajennettiin johdosperheellä eli muodostettiin osien yhdistelmäkysely (T2/ABCabc), tämän saanti nousi niin JA- kuin virkeoperaattoriakin käytettäessä aivan samaksi kuin perinteisen osien yhdistelmäkyselyn.

Kun T2-tutkimusympäristön yhdistelmäkyselyn (ABC) **tarkkuutta** verrataan saman tutkimusympäristön osien yhdistelmäkyselyn (ABCabc) tarkkuuteen, niin JA-operaattoria käytettäessä ensinmainitun kyselytyypin tarkkuusarvo oli 35 prosenttiyksikköä korkeampi kuin jälkimmäisen. Virkeoperaattoria käytettäessä yhdistelmäkyselyn tarkkuusarvo taas oli reilun yhden prosenttiyksikön verran **matalampi** kuin osien yhdistelmäkyselyn tarkkuus.

T1- ja T2-ympäristöjen kaikki kyselyt, joissa yhdyssanat olivat kokonaisia (T1/ABC, T2/AC, T2/ABC), olivat tarkkuudeltaan parempia kuin T1- ja T2-ympäristöjen osien yhdistelmäkyselyt (ABCabc), kun hakusanat oli kytketty toisiinsa JA-operaattorilla. Erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.01 ja Sparck Jonesin mittarin mukaan käytännössä olennaisia.

Osien yhdyssanakyselyn (ACac) osalta erot eivät olleet yhtä selvät kuin edellä: kun hakusanat oli yhdistetty JA-operaattorilla, T2-ympäristön osien

yhdyssanakyselyn tarkkuus oli noin 30 prosenttiyksikköä yhdyssana- (AC) ja yhdistelmäkyselyjen (ABC) tarkkuutta alempi. Erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.05 ja käytännössä olennaisia. - Toisaalta osien yhdyssanakyselyn tarkkuus oli JA-operaattoria käytettäessä 5 prosenttiyksikköä korkeampi kuin T1-ympäristön osien yhdistelmäkyselyn, ja ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.05.

Virkeoperaattoria käytettäessä osien yhdyssanakyselyn tarkkuus oli 3 prosenttiyksikköä **huonompi** kuin perinteisen osien yhdistelmäkyselyn (T1/ABCabc). Tällöin eri kyselytyyppien tarkkuusarvojen välillä ei ollut lainkaan tilastollisesti merkitseviä eroja.

Kun osien yhdyssanakyselyä laajennettiin johdosperheellä eli T2-ympäristössä muodostettiin osien yhdistelmäkysely (ABCabc), tämän tarkkuus oli JA-operaattoria käytettäessä vajaan prosenttiyksikön verran parempi ja virkeoperaattoria käytettäessä jokseenkin sama kuin T1-ympäristön osien yhdistelmäkyselyn tarkkuus.

Yhdyssanaosajoukossa hakusanojen osittaminen siis on hyvin tehokas keino parantaa saantia. JA-operaattoria käytettäessä tarkkuus tosin alenee vastaavasti kuin saanti kasvaa, mutta virkeoperaattoria käytettäessä eri kyselytyyppien tarkkuusarvojen välillä ei ollut sen enempää tilastollisesti kuin käytännössäkään merkitseviä eroja - tiukka läheisyysoperaattori karsii väärät osumat tehokkaasti pois.

Virkeoperaattoria käytettäessä yhdyssanat siis kannattaa aina jakaa osiinsa, tarkkuus ei juuri heikkene (tällä aineistolla jopa lievästi parantui) ja saantiarvojen väliset erot ovat selvästi tilastollisesti merkitseviä (merkitsevyystasolla 0.01) ja myös käytännössä olennaiset.

9.2.2 Finstems- ja Hahmotin-ohjelmien väliset erot

Vakiokyselyissä Finstems- ja Hahmotin-ohjelmien tuottamat vartalot olivat melko samanlaisia. Finstems-ohjelmalla oli taipumus tuottaa substantiiveista pidempiä vartaloita kuin Hahmotin-ohjelmalla, kun taas jälkimmäisen ohjelman tuottamat verbivartalot olivat pidempiä. Esimerkiksi kauppa-substantiivin monikkovartalot olivat Finstems-ohjelmalla kaupoi, kauppoi ja kauppoj ja Hahmotin-ohjelmalla taas kaupoi ja kauppo. Hahmotin tuotti rakentaa-verbistä vartalot rakenna, rakenta, rakensi, rakennet, rakenni ja rakenti, kun Finstems tuotti vain vartalot rakent, rakenn ja rakensi. Seitse-

mässä kyselyssä nämä taivutusvartaloiden erot vaikuttivat myös tulosjoukon kokoon.

Yhdessä tapauksessa (kyselyssä 17) Finstems-ohjelman substantiivivartaloita laina, lainoi ja lainoj hakusanoina käyttäen tietokannasta saatiin 68 dokumenttia, kun taas Hahmotin-ohjelman lyhyemmät vartalot laina ja laino tuottivat 69 dokumenttia. Näin löydetty yksi lisädokumentti ei ollut kyselyn kannalta relevantti.

Viidessä kyselyssä (kyselyt 8, 10, 14, 20 ja 29) Hahmotin-ohjelmalla saadut tulosjoukot olivat pienempiä kuin Finstems-ohjelmalla saadut, koska pidemmät verbivartalot karsivat osan dokumenteista pois. Näistä viidestä kyselystä neljässä Finstems-ohjelman tulosjoukko sisälsi vain yhden dokumentin enemmän kuin Hahmotin-ohjelman tulosjoukko. Lisädokumenteista yksikään ei ollut relevantti.

Yhdessä tapauksessa (kysely 8: kirkkojen rakentaminen) eri vartalo-ohjelmilla saatujen tulosjoukkojen ero oli suurempi. Hahmotin tuotti rakentaa-verbistä vartalot rakenna, rakenta, rakensi, rakennet, rakenni ja rakenti. Näitä vartaloita hakusanoina käytettäessä saatiin 126 dokumenttia. Finstems tuotti rakentaa-verbistä lyhyemmät vartalot rakent, rakenn ja rakensi. Näitä vartaloita käytettäessä saatiin 145 dokumenttia. Finstems-ohjelman vartaloilla saadut ylimääräiset sananmuodot olivat enimmäkseen rakenne- ja rakenteellinen-sanojen esiintymiä (12 kpl); myös muutama rakennuttaa-, rakentua-, rakennelma- ja rakennella-sanan taivutusmuoto oli tullut mukaan. Finstems-ohjelmaa käytettäessä saaduista 19 lisädokumentista vain yksi oli arvioijien mielestä jossain määrin relevantti. Siinäkään aiheena ei ollut uudisrakentaminen, vaan kirkkojen korjaus (kirkon rakenne).

Vakiokyselyissä Finstems- ja Hahmotin-ohjelmien tuottamat vartalot eivätkä siten niiden tuottamat tulosjoukot kovin paljon poikenneet toisistaan. Mikäli vartalot poikkesivat toisistaan niin paljon, että eri ohjelmien vartaloilla saadut tulosjoukot olivat erikokoisia, lisädokumentit eivät tämän tutkimuksen kyselyissä olleet relevantteja kuin yhdessä tapauksessa. Tämän perusteella ei siis ollut suurta eroa, katkaistiinko vakiokyselyjen hakusanat Finstems- vai Hahmotin-ohjelmalla. Niinpä tämän tutkimuksen merkitsevyystestit laskettiin T2-ympäristössä vain Finstems-ohjelman tuottamien hakutulosten perusteella.

9.3 Hakijan katkaisemien ja seulottujen hakusanojen vertailu

T3-ympäristön vertailu erosi muista tutkimusympäristöistä siinä, että hakupyynnön 28 tulosjoukko oli virkeoperaattoria käytettäessä tyhjä, jolloin sen tarkkuusarvoa ei voitu laskea. Niinpä tämä kysely jätettiin vertailuista pois. Siten T3-ympäristön perusjoukossa on 25 kyselyä ja vertailut sen ja T1-ympäristön kanssa tehtiin näitä 25 kyselyä käyttäen.

9.3.1 Perusjoukko

Yhdyssanakysely

Taulukosta 20 näkyy, että T3-ympäristön yhdyssanakyselyn (AC) tulosjoukot olivat **kooltaan** selvästi pienempiä kuin vastaavat perinteisen yhdistelmäkyselyn (ABC) tulosjoukot. Tämä päti erityisesti JA-operaattorilla haettaessa, mutta myös virkeoperaattoria käytettäessä. Tämän lisäksi myös T3-tutkimusympäristön yhdistelmäkyselyjen (ABC) tulosjoukot olivat pienempiä kuin vastaavat perinteisen yhdistelmäkyselyn tulosjoukot.

Seulottaessa **tarkkuusarvot** olivat korkeampia kuin perinteisellä tavalla haettaessa. Seulonnan yhdyssanakyselyn (AC) tulosjoukko sai sekä JA- että virkeoperaattoria käytettäessä noin 4 prosenttiyksikköä korkeamman tarkkuusarvon kuin perinteisen yhdistelmäkyselyn (ABC) tulosjoukko.

Tarkkuusarvojen väliset erot olivat JA-operaattoria käytettäessä tilastollisesti merkitseviä merkitsevyystasolla 0.05, joten sen osalta hypoteesi 2 sai lievästi tukea. Erot eivät kuitenkaan olleet käytännössä merkitseviä. Virke-

Taulukko 20. T1- ja T3-ympäristöissä kyselytyypeillä AC ja ABC saatujen hakutulosten väliset erot perusjoukossa (N =25).

Operaattori	Kyselytyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T3**	**T1**	**T3**	**T1**	**T3**
JA	AC		11,2		63,5		72,8
	ABC	**20,6**	**16,9**	**73,8**	**70,0**	**68,4**	**69,8**
Virke	AC		7,8		54,0		80,7
	ABC	**11,6**	**10,2**	**61,8**	**59,1**	**77,0**	**78,4**

AC Yhdyssanakysely
ABC Yhdistelmäkysely
T1 Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T3 Taivutusmuotohakemisto, seulotut hakusanat

operaattoria käytettäessä taas eri kyselytyyppien väliset erot eivät olleet tilastollisesti eivätkä siten myöskään käytännössä merkitseviä. Se, että tarkkuusarvot virkeoperaattoria käytettäessä ovat eri kyselytyypeillä niin lähellä toisiaan selittynee sillä, että virkeoperaattori rajaa kyselyä muutenkin niin paljon, ettei hakusanojen parempi tarkkuus enää vaikuta kokonaistulokseen niin paljon kuin JA-operaattorilla.

Seulonnan suhteelliset **saantiarvot** olivat huonommat kuin hakijan katkaisemia hakusanoja käytettäessä. Yhdyssanakyselyn (AC) tulosjoukon saantiarvo oli JA-operaattoria käytettäessä 10 prosenttiyksikköä alempi ja virkeoperaattorilla noin 8 prosenttiyksikköä alempi kuin T1-ympäristön yhdistelmäkyselyn (ABC) tulosjoukon saantiarvo. Näiden kyselytyyppien saantiarvojen väliset erot olivat sekä JA- että virkeoperaattorin tapauksessa tilastollisesti merkitseviä merkitsevyystasolla 0.01. Sparck Jonesin mittarin perusteella ero oli JA-operaattorin tapauksessa käytännössä olennainen, virkeoperaattoria käytettäessä ero oli huomattava. Tämä tulos oli täysin hypoteesin 2 mukainen.

Yhdistelmäkysely

Kun seulonnassa otettiin myös johdokset mukaan kyselyyn, tämän yhdistelmäkyselyn (ABC) tulosjoukon **tarkkuusarvo** oli sekä JA- että virkeoperaattoria käytettäessä reilun yhden prosenttiyksikön verran korkeampi kuin T1-ympäristössä saadun yhdistelmäkyselyn tulosjoukon tarkkuusarvo. Tarkkuusarvojen väliset erot eivät olleet tilastollisesti merkitseviä, joten hypoteesi 3 ei siis saanut tukea.

Seulonnan yhdistelmäkyselyn **saanti** jäi JA-operaattoria käytettäessä noin 4 prosenttiyksikköä alemmaksi kuin perinteisen yhdistelmäkyselyn saanti. Tulos oli sama virkeoperaattoria käytettäessä: T3-ympäristön yhdistelmäkyselyn saanti jäi 3 prosenttiyksikköä alemmaksi kuin T1-ympäristön yhdistelmäkyselyn. Kummassakaan tapauksessa saantiarvojen väliset erot eivät olleet tilastollisesti merkitseviä. Tämä on periaatteessa linjassa hypoteesin 3 kanssa, mutta hypoteesin mukaan saannin olisi pitänyt olla sama – nyt seulonnan saanti jäi lievästi perinteistä huonommaksi. Koska ero ei kuitenkaan ollut suuri, tilastollisesti merkitseviä eroja ei syntynyt.

9.3.2 Johdososajoukko

Yhdyssanakysely

Kun T3-ympäristössä tarkastellaan erikseen niitä kahdeksaa hakupyyntöä, joissa hakusanoilla oli aitoja johdoksia (taulukko 21; liite 14, luku 2), yhdyssanakyselyn (AC) **tarkkuusarvot** olivat 11 prosenttiyksikköä perinteistä yhdistelmäkyselyä (ABC) korkeammat sekä JA- että virkeoperaattoria käytettäessä. Ero ei virkeoperaattorin tapauksessa ollut tilastollisesti merkitsevä, mutta JA-operaattoria käytettäessä tarkkuusarvojen väliset erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.05. JA-operaattorin tapauksessa ero oli käytännössä olennainen. Sen osalta tulos tukee hypoteesia 2.

T3-ympäristön yhdyssanakyselyn (AC) **saantiarvot** olivat tässä osajoukossa selvästi perinteisen yhdistelmäkyselyn (ABC) saantiarvoja huonommat: JA-operaattoria käytettäessä 29 ja virkeoperaattoria käytettäessä 23 prosenttiyksikköä alemmat. Molemmissa tapauksissa T1- ja T3- ympäristöjen saantiarvojen välinen ero oli merkitsevä vielä merkitsevyystasolla 0.01. Erot olivat käytännössäkin olennaiset. Tämä tulos siis tuki vahvasti hypoteesin 2 väitettä saannin alenemisesta.

Seulonnan johdososajoukon yhdyssanakyselyssä saanti siis aleni olennaisesti enemmän kuin tarkkuus parani, kun vertauskohteena oli kysely, jossa käytettiin hakijan itse katkaisemia hakusanoja. Vaikka tulos tukeekin hypoteesia 2, käytännön tavoite ei kuitenkaan toteudu eli tarkkuuden parantuminen laskee saantia liikaa.

Taulukko 21. T1- ja T3-ympäristöissä kyselytyypeillä AC ja ABC saatujen hakutulosten väliset erot johdososajoukossa (N = 8).

Operaattori	Kyselytyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T3**	**T1**	**T3**	**T1**	**T3**
JA	AC		13,1		53,4		57,7
	ABC	**41,3**	**31,0**	**82,7**	**73,7**	**46,9**	**48,3**
Virke	AC		4,8		33,7		75,2
	ABC	**15,9**	**12,3**	**56,2**	**49,9**	**64,3**	**67,9**

AC Yhdyssanakysely
ABC Yhdistelmäkysely
T1 Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T3 Taivutusmuotohakemisto, seulotut hakusanat

Kun tutkimusympäristön T3 kyselyt laajennettiin yhdistelmäkyselyiksi (ABC) lisäämällä niihin johdosperhe, seulonnan **tarkkuusarvot** olivat reilun prosenttiyksikön (JA-operaattoria käytettäessä) tai vajaat 4 prosenttiyksikköä (virkeoperaattori) korkeammat kuin T1-ympäristön yhdistelmäkyselyn. Erot eivät olleet tilastollisesti merkitseviä. Tämä ei ollut hypoteesin 3 mukaista, vaan tarkkuuden olisi pitänyt olla selvästi parempi.

Seulonnan **saanti** puolestaan jäi yhdistelmäkyselyssä (ABC) JA-operaattoria käytettäessä 9 ja virkeoperaattoria käytettäessä reilut 6 prosenttiyksikköä alemmaksi kuin perinteisen yhdistelmäkyselyn. Nämä erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.05 ja myös käytännössä huomattavia.

Hypoteesin 3 mukaan T3-ympäristön yhdistelmäkyselyn saannin olisi pitänyt olla sama ja tarkkuuden parempi kuin T1-ympäristön yhdistelmäkyselyn, joten tulos ei ole hypoteesin 3 mukainen sen enempää saannin kuin tarkkuudenkaan osalta

Seulonnan voi sanoa olevan automaattisen katkaisun jatkojalostusta, jossa tavoitteena on parantaa tarkkuutta ja jättää saanti ennalleen. Johdososajoukossa seulonta ei kuitenkaan erityisemmin kohentanut haun tarkkuutta verrattuna perinteiseen yhdistelmäkyselyyn; sen sijaan saanti romahti perinteiseen yhdistelmäkyselyyn verrattuna, kun johdokset jäivät kyselyistä pois. Johdosten lisäys yhdistelmäkyselyyn ei lopulta nostanut saantia niin korkealle kuin oli tarkoitus – tarkkuudenhan olisi pitänyt parantua niin, että saanti ei olisi kärsinyt lainkaan.

9.3.3 Yhdyssanaosajoukko

Kun tutkitaan sitä yhdeksän kyselyn osajoukkoa, jossa yhdyssanat voitiin jakaa osiinsa (taulukko 22; liite 14, luku 3), nähdään samalla tavoin kuin T2-ympäristössä, että hakutulokset muuttuvat huomattavasti, kun yhdyssanat jaetaan osiinsa ja osat lisätään kyselyyn mukaan.

Kun esimerkiksi yhdistelmäkysely (ABC) laajennettiin T3-tutkimusympäristössä osien yhdistelmäkyselyksi (ABCabc), hakutuloksen **saanti** kasvoi JA-operaattoria käytettäessä 29 prosenttiyksikköä ja virkeoperaattoria käytettäessä 14 prosenttiyksikköä. Sparck Jonesin mittarin mukaan vertailtujen vaihtoehtojen väliset erot olivat käytännössä olennaiset. - Samalla **tarkkuus**

JA-operaattorin tapauksessa huononi 27 prosenttiyksikköä ja virkeoperaattoria käytettäessä parani reilun prosenttiyksikön verran.

Kaikki ne kyselytyypit, joissa yhdyssanoja ei ollut ositettu (T3-ympäristön yhdyssanakysely AC ja molempien ympäristöjen yhdistelmäkyselyt ABC), erosivat JA-operaattoria käytettäessä saanniltaan merkitsevyystasolla 0.01 kaikista ositetuista kyselytyypeistä (siis T3-ympäristön osien yhdyssanakyselystä ACac ja molempien tutkimusympäristöjen osien yhdistelmäkyselystä ABCabc.)

Muuten vastaava tulos saatiin virkeoperaattoria käytettäessä, paitsi että T3-ympäristön osien yhdyssanakyselyn (ACac) ja T1-ympäristön yhdistelmäkyselyn (ABC) saantiarvojen välinen ero oli merkitsevä merkitsevyystasolla 0.05 – muuten ositettuja ja osittamattomia yhdyssanoja sisältäneiden tulosjoukkojen saantiarvojen väliset erot olivat merkitseviä merkitsevyystasolla 0.01.

Osien yhdyssanakyselyn (T3/ACac) saanti oli JA-operaattoria käytettäessä reilut 11 ja virkeoperaattoria käytettäessä vajaat 4 prosenttiyksikköä alempi kuin perinteisen osien yhdistelmäkyselyn (ABCabc). Saantiarvojen välinen ero oli JA-operaattorilla tilastollisesti merkitsevä merkitsevyystasolla 0.01

Taulukko 22. T1- ja T3-ympäristöissä kyselytyypeillä ACac ja ABCabc saatujen hakutulosten väliset erot yhdyssanaosajoukossa (N = 9).

Operaattori	Kyselytyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T3**	**T1**	**T3**	**T1**	**T3**
JA	AC		6,8		52,8		77,9
	ABC	8,0	7,2	55,7	54,5	76,1	77,7
	ACac		15,2		76,8		53,8
	ABCabc	**25,8**	**18,2**	**88,1**	**83,5**	**42,1**	**50,4**
Virke	AC		5,7		46,5		76,4
	ABC	6,1	6,0	48,8	48,3	80,0	79,8
	ACac		7,8		59,3		77,3
	ABCabc	**8,6**	**8,3**	**62,8**	**62,2**	**81,1**	**81,2**

AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
ACac	Osien yhdyssanakysely
ABCabc	Osien yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T3	Taivutusmuotohakemisto, seulotut hakusanat

ja käytännössä olennainen (Sparck Jonesin mittari). JA-operaattorin tulos oli hypoteesin 2 mukainen.

Lisäksi osien yhdyssanakyselyn saanti erosi saman T3-ympäristön osien yhdistelmäkyselyn saannista tilastollisesti merkitsevästi merkitsevyystasolla 0.05, kun hakusanat oli yhdistetty JA-operaattorilla. Tässä tapauksessa ero oli käytännössä huomattava. Tämäkin tukee hypoteesia 2.

Kun johdokset lisättiin kyselyyn mukaan eli T3-ympäristössä muodostettiin osien yhdistelmäkysely (ABCabc), seulonnan saantiarvot olivat JA-operaattoria käytettäessä vielä vajaat 5 prosenttiyksikköä ja virkeoperaattoria käytettäessä alle prosenttiyksikön alemmat kuin perinteisen osien yhdistelmäkyselyn. Näiden kyselytyyppien saantiarvojen välinen ero ei kuitenkaan ollut tilastollisesti merkitsevä. Näin ollen tulos ei ole suoranaisesti hypoteesin 3 vastainen, joskin siinä saannin väitettiin olevan saman.

Kun T3-tutkimusympäristön yhdistelmäkyselyn (ABC) **tarkkuutta** verrataan T1-ympäristön osien yhdistelmäkyselyn (ABCabc) tarkkuuteen, niin JA-operaattoria käytettäessä ensinmainitun kyselytyypin tarkkuusarvo oli lähes 36 prosenttiyksikköä korkeampi kuin jälkimmäisen. Virkeoperaattoria käytettäessä seulonnan yhdistelmäkyselyn (ABC) tarkkuusarvo oli jopa reilun yhden prosenttiyksikön verran **matalampi** kuin T1-ympäristön osien yhdistelmäkyselyn (ABCabc) tarkkuus.

Eli seulonnassa saattoi käydä niin, että relevantteja dokumentteja sen tuloksena jäi tulosjoukosta pois niin paljon, että tulosjoukkoon itse asiassa jäi suurempi osuus epärelevantteja dokumentteja kuin perinteisellä hakutavalla saadussa tulosjoukossa.

Kun hakusanat oli kytketty JA-operaattorilla, oli T1-ympäristön osien yhdistelmäkysely (ABCabc) tarkkuudeltaan huonoin (taulukko 22). Lisäksi sen tarkkuusarvo Friedmanin merkitsevyystestin mukaan erosi kaikkien muiden kyselytyyppien tarkkuusarvoista tilastollisesti merkitsevästi tasolla 0.01. Sen sijaan muiden kyselytyyppien tarkkuusarvojen välillä keskenään ei ollut tilastollisesti merkitseviä eroja JA-operaattoria käytettäessä. (Paitsi T3-ympäristön yhdyssanakyselyn ja osien yhdistelmäkyselyn välillä; tosin tässäkään tapauksessa tarkkuusarvojen välisillä eroilla ei käytännössä ollut merkitystä, koska erot olivat tilastollisesti merkitsevät vasta merkitsevyystasolla 0.1).

Virkeoperaattoria käytettäessä eri kyselytyyppien tarkkuusarvojen välille ei löytynyt tilastollisesti merkitseviä eroja - ei edes silloin, kun vertailukohteena oli T1-ympäristön osien yhdistelmäkysely. Virkeoperaattori siis rajaa kyselyn alaa jo niin tarkasti, että seulonta ei enää tulosta paranna.

T3-ympäristön osien yhdistelmäkyselyn (ABCabc) tarkkuus oli JA-operaattoria käytettäessä yli 8 prosenttiyksikköä parempi kuin perinteisen osien yhdistelmäkyselyn. T1- ja T3-ympäristöjen tarkkuusarvojen välinen ero oli tässä tapauksessa tilastollisesti merkitsevä merkitsevyystasolla 0.01. Sparck Jonesin määrittelyn mukaan tarkkuusarvojen välinen ero oli JA-operaattorin tapauksessa käytännössä huomattava. Tämä oli hypoteesin 3 mukaista.

9.3.4 Seulonnan toteutustavasta

FULLTEXT-projektin testauksissa seulontaan kuluvaa aikaa ei mitenkään optimoitu. Niinpä seulonta kesti kauan erityisesti tapauksissa, joissa sitä käytännössä eniten tarvittaisiin eli lyhyiden hakusanojen tarkistamisessa: lyhyeksi katkaistut hakusanat palauttavat paljon hakemistosanoja, joiden kaikkien käsittely vie paljon aikaa. Testiajoissa esimerkiksi maa- ja mai-vartaloilla saatuja hakemistosanoja ei useista yrityksistä huolimatta pystytty lainkaan seulomaan, koska käyttäjän ja hakujärjestelmän välinen linjaliikenneyhteys katkesi aina ennen kuin aliohjelma oli ehtinyt tarkistaa osumat.

Testikyselyissä *maa**-hakusanalla saatiin 9 831 osumaa ja *maa**- ja *mai**-hakusanoilla yhteensä 11 076 osumaa. Tietokannassa oli 23 244 artikkelia, joista kertyi taivutusmuotohakemistoon kaikkiaan 516 371 erilaista hakemiston merkkijonoa. Realistisen kokoiset tuotantotietokannat ovat kooltaan ainakin kymmenkertaisia, paremminkin sata- tai tuhatkertaisia tähän testitietokantaan verrattuna. Koska seulottavien hakemistosanojen määrä oletettavasti kasvaa samassa suhteessa kuin tietokantakin kasvaa, seulonta käytännössä kestää sitä kauemmin, mitä enemmän sitä tarvittaisiin - olkoonkin, että tulkintaohjelmien tarvitsema aika olisi optimoitu minimiin. Jos hakujärjestelmän ylläpitäjä hinnoittelee palvelunsa toimintojen mukaan, esimerkiksi hinnoittamalla komennot niiden kuluttamien tietokoneresurssien mukaisesti, voi käydä niin, että hakijat eivät käytäkään automaattista seulontaa, koska se tulee kalliiksi suhteessa siitä saataviin hyötyihin.

Tutkimusympäristön T3 tulosjoukkojen koon supistuminen verrattuna tutkimusympäristön T2 tulosjoukkojen kokoon todistaa osaltaan, että katkaistut hakusanat tuottavat tulosjoukkoon myös sanoja, jotka eivät ole hakusanan

esiintymiä, vaan vain sattuvat alkamaan samalla tavalla. Kun näin saadut epärelevantit dokumentit seulotaan tulosjoukosta pois, tulosjoukon koko selvästi pienenee.

Vaikka hakusanojen seulonta paransikin hakujen tarkkuutta, se tapahtui korostetusti saannin kustannuksella, koska hakemistosanojen oli täsmättävä tarkalleen hakusanoihin. Näin karsiutuivat muun muassa hakusanalla alkavat yhdyssanat ja sanat, joita perusmuoto-ohjelma ei pystynyt tunnistamaan. (Mikäli yhdyssanan yksikin osa on perusmuoto-ohjelmalle tuntematon, jää koko yhdyssana tunnistamatta.)

Seulontaperiaatteita on mahdollista lieventää FULLTEXT-projektissa käytetystä menetelmästä: saannin parantamiseksi tulosjoukkoihin voidaan sallia tulevaksi myös ne dokumentit, joissa esiintyy hakusanoilla alkavia yhdyssanoja ja tuntemattomia sananmuotoja. Näiden lievennysten seurauksena tarkkuus kuitenkin heikkenee. Tämän tutkimuksen tiukka seulontamenetelmäkään ei nostanut tarkkuusarvoja selvästi perinteisen hakutavan tarkkuusarvoja korkeammiksi muualla kuin yhdyssanaosajoukossa. Jos seulonta toteutetaan väljemmin, lopputulos ei enää välttämättä poikkea paljoakaan seulomattoman haun tuloksesta, joten seulonnan hyöty jää olemattomaksi. Tarkkuuttahan voidaan parantaa muillakin, ei-lingvistisillä keinoilla, vaikkapa käyttämällä tiukempaa läheisyysoperaattoria eli soveltamalla perinteisiä tiedonhakujärjestelmien keinoja.

9.4 Taivutusmuoto- ja perusmuotohakemistosta saatujen hakutulosten vertailu

Tutkimusympäristössä T4 hakija antoi hakusanat perusmuodossa. Peruskyselyssä (A) hakija syötti pelkästään hakusanojen perusmuodot, johdoskyselyssä (AB) lisäksi hakusanan johdosperheen jäsenet perusmuodossa. Yhdyssanakysely (AC) toteutettiin lisäämällä peruskyselyyn hakusanan rinnalle vartalot, jotka vartalo-ohjelma oli tuottanut - siis hakusanojen automaattinen katkaisu samalla tavalla kuin tutkimusympäristössä T2. Näillä vartalo-ohjelman tuottamilla vartaloilla haettiin sekä perusmuotohakemistosta että tunnistamattomien sanojen hakemistosta. Myös johdoskysely (AB) laajennettiin vastaavalla tavalla yhdistelmäkyselyksi (ABC) eli hakusanan johdosperheen jäsenistä tuotettiin vartalo-ohjelmalla vartalot, jotka sitten lisättiin kyselyyn alkuperäisen hakusanan rinnalle.

9.4.1 Perusjoukko

Peruskysely

Pelkillä hakusanan perusmuodoilla hakeminen pienensi tulosjoukon keskimääräisen **koon** noin puoleen verrattuna perinteisellä yhdistelmäkyselyllä (ABC) saadun tulosjoukon kokoon (taulukko 23).

T4-ympäristön peruskyselyn (A) **saantiarvot** olivat selvästi huonommat kuin perinteisen yhdistelmäkyselyn: JA-operaattoria käytettäessä saanti laski 17 prosenttiyksikköä ja virkeoperaattoria käytettäessä 14 prosenttiyksikköä alemmas kuin perinteisen yhdistelmäkyselyn saanti.

Sekä JA-operaattoria että virkeoperaattoria käytettäessä saantiarvojen välinen ero oli Friedmanin testin perusteella tilastollisesti merkitsevä tasolla 0.01. Molemmissa tapauksissa ero oli käytännössä olennainen. Tämä tulos oli hypoteesin 4 mukainen. (Liite 14, luku 1.)

Peruskyselyn keskimääräinen **tarkkuus** oli perinteisen yhdistelmäkyselyn tarkkuutta korkeampi, noin 8 prosenttiyksikköä sekä JA- että virkeoperaattorilla. Tarkkuusarvojen välinen ero oli JA-operaattoria käytettäessä tilastollisesti merkitsevä merkitsevyystasolla 0.01 ja Sparck Jonesin mittarin mukaan käytännössä huomattava. Sen sijaan virkeoperaattoria käytettäessä tarkkuusarvojen välinen ero ei ollut tilastollisesti merkitsevä.

Peruskyselyn tarkkuus ei siis noussut yhtä monta prosenttiyksikköä kuin saanti laski, kun vertailukohteena oli perinteinen yhdistelmäkysely (ABC).

JA-operaattoria käytettäessä hypoteesi 4 siis piti paikkansa, virkeoperaattorilla ei tarkkuuden osalta – virkeoperaattoria käytettäessä tarkkuus on muutenkin korkea, joten tarkkuuden parantaminen vaatisi suhteessa enemmän ponnistuksia kuin JA-operaattoria käytettäessä.

Johdoskysely

Kun peruskyselyä laajennettiin johdoksilla (kyselytyyppi AB), **saantiarvot** paranivat jonkin verran verrattuna pelkkään peruskyselyyn. Johdoskyselyn saantiarvot jäivät silti alle perinteisen yhdistelmäkyselyn (ABC) saantiarvojen. JA-operaattoria käytettäessä T4-ympäristön johdoskyselyn (AB) saanti oli reilut 10 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn. Virkeoperaattoria käytettäessä saanti taas jäi 8 prosenttiyksikköä alemmaksi kuin perinteisessä yhdistelmäkyselyssä. Kummassakin tapauksessa saanti-

Taulukko 23. T1- ja T4-ympäristöissä kyselytyypeillä A, AB, AC ja ABC saatujen hakutulosten väliset erot perusjoukossa (N = 26).

Ope-raattori	Kysely-tyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T4**	**T1**	**T4**	**T1**	**T4**
JA	A		8,9		54,7		75,8
	AB		13,5		61,7		72,4
	AC		12,4		64,1		73,0
	ABC	**20,0**	**19,7**	**72,0**	**72,0**	**68,0**	**70,3**
Virke	A		6,0		46,3		84,2
	AB		8,1		52,0		80,4
	AC		8,3		54,4		79,2
	ABC	**11,3**	**11,2**	**60,1**	**60,2**	**76,6**	**77,3**

A	Peruskysely
AB	Johdoskysely
AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat

arvojen väliset erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.01. Sparck Jonesin mittarin mukaan saantiarvojen välinen ero oli käytännössä JA-operaattorilla olennainen ja virkeoperaattorilla huomattava.

Tulos ei ole lainkaan hypoteesin 5 mukainen, koska sen mukaan saannin piti olla keskimäärin sama. Pelkkien johdosten lisääminen kyselyyn ei siis riitä nostamaan saantiarvoja niin paljon, että peruskyselyn ja johdoskyselyn saantiarvojen välille olisi syntynyt tilastollisesti merkitseviä eroja.

Toisaalta johdosten lisäys ei alentanut tarkkuusarvojakaan kuin hiukan verrattuna peruskyselyn tarkkuusarvoihin: T4-ympäristön johdoskyselyn **tarkkuus** oli sekä JA-että virkeoperaattoria käytettäessä 4 prosenttiyksikön verran korkeampi kuin perinteisen yhdistelmäkyselyn. Tarkkuusarvojen väliset erot eivät olleet tilastollisesti merkitseviä. Tässäkään tapauksessa tulos ei ole aivan hypoteesin 5 mukainen, koska tarkkuuden piti olla parempi kuin perinteisen yhdistelmäkyselyn. Nyt saanti aleneni useampia prosenttiyksikköjä enemmän kuin tarkkuus parani.

Toisaalta on todettava, että kun peruskysely laajennettiin johdoskyselyksi, saantiarvot paranivat JA-operaattoria käyttäessä 7 ja virkeoperaattoria käytettäessä vajaat 6 prosenttiyksikköä. JA-operaattorin tapauksessa ero oli ti-

lastollisesti merkitsevä merkitsevyystasolla 0.05, virkeoperaattorilla merkitsevyystasolla 0.01. Ero oli Sparck Jonesin mittarin perusteella käytännössä huomattava. Vastaavasti johdoskyselyn tarkkuus jäi sekä JA- että virkeoperaattorilla vain reilut 3 prosenttiyksikköä alemmaksi kuin peruskyselyn. Virkeoperaattorin tapauksessa ero ei ollut tilastollisesti merkitsevä eikä varsinaisesti JA-operaattorillakaan (ero oli merkitsevä vasta merkitsevyystasolla 0.1 eli käytännössä olematon). Johdokset lisäämällä saantia siis pystytään parantamaan ilman, että tarkkuus vastaavasti huononisi.

Yhdyssanakysely

Kun hakusanan perusmuodon lisäksi haettiin myös hakusanalla alkavat yhdyssanat (yhdyssanakysely AC), suhteellinen **saanti** oli JA-operaattoria käytettäessä 8 ja virkeoperaattoria käytettäessä vajaat 6 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn (ABC). Saantiarvojen väliset erot olivat sekä JA- että virkeoperaattoria käytettäessä tilastollisesti merkitseviä merkitsevyystasolla 0.01. Sparck Jonesin mittarin mukaan erot olivat käytännössä huomattavat. Pelkästään yhdyssanojen lisäys peruskyselyyn ei siis kasvata saantia niin paljon, että saanti nousisi perinteisen yhdistelmäkyselyn lukemiin.

Yhdyssanakyselyn **tarkkuus** puolestaan oli JA-operaattoria käytettäessä 5 ja virkeoperaattoria käytettäessä vajaat 3 prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn. Eri tutkimusympäristöjen tarkkuusarvojen välinen ero oli JA-operaattorilla tilastollisesti merkitsevä merkitsevyystasolla 0.01, sen sijaan virkeoperaattoria käytettäessä ero ei ollut merkitsevä. Ero oli JA-operaattorin tapauksessa juuri ja juuri käytännössä huomattava. Yhdyssanakyselyn tarkkuus siis on kohtuullinen, joskin sen pitäisi selvemmin olla perinteistä yhdistelmäkyselyä parempi.

Yhdyssanakysely voitaisiin T4-ympäristössä tuottaa peruskyselystä (A) automaattisesti vartalo-ohjelman avulla siten, että hakija vain syöttää järjestelmälle hakusanan perusmuodot, joista tuotetaan vartalot ja sitten suoritetaan haku näillä vartaloilla. Yhdyssanakyselyn (AC) **saanti** oli JA-operaattoria käytettäessä reilut 9 ja virkeoperaattoria käytettäessä 8 prosenttiyksikköä korkeampi kuin peruskyselyn (A) saanti. Kummassakin tapauksessa ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01. Ero oli Sparck Jonesin mittarin perusteella käytännössä huomattava.

Yhdyssanakyselyn **tarkkuus** taas jäi JA-operaattorilla vain 3 ja virkeoperaattorilla 5 prosenttiyksikköä alemmaksi kuin pelkän peruskyselyn. Kummassakaan tapauksessa ero ei ollut tilastollisesti merkitsevä.

Edellä kuvatun perusteella voidaan sanoa, että T4-ympäristön kyselyissä ei kannata käyttää pelkkiä hakusanojen perusmuotoja, vaan kyselyt pitäisi samantien laajentaa yhdyssanakyselyiksi. Näin hakutulosten saantia voidaan parantaa huomattavasti verrattuna pelkkään peruskyselyyn ilman, että hakutulosten tarkkuus vastaavasti huononisi.

Yhdistelmäkysely

Kun kyselyt T4-ympäristössä laajennettiin yhdistelmäkyselyiksi (ABC), T1- ja T4-ympäristöjen tulosjoukkojen **koot** olivat jokseenkin samat.

T4-ympäristössä **saantiarvot** olivat JA-operaattoria käytettäessä täsmälleen samat kuin T1-ympäristössä ja eikä eroja virkeoperaattoriakaan käytettäessä ollut käytännöllisesti katsoen lainkaan. Tämä tulos ei vastaa hypoteesia 6, jonka mukaan T4-ympäristön yhdistelmäkyselyn saannin pitäisi olla paremman kuin vastaavan perinteisen kyselytyypin saannin.

Kun T4-ympäristön yhdistelmäkyselyn tulosjoukon saantiarvoja verrattiin saman T4-ympäristön johdoskyselyn (AB) ja yhdyssanakyselyn (AC) saantiarvojen kanssa, erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.01. Eli peruskyselyn laajentaminen pelkästään johdosperheellä tai pelkästään yhdyssanoilla ei riitä nostamaan tulosjoukon saantiarvoa perinteisen yhdistelmäkyselyn saantiarvon tasalle; tarvitaan nuo molemmat laajennukset, jotta T4-ympäristön tulosjoukkojen saanti ei jäisi alle perinteisen yhdistelmäkyselyn saannin. Tällöinkään saantiarvo ei siis nouse korkeammaksi kuin perinteisen yhdistelmäkyselyn saantiarvo.

Tarkkuusarvot puolestaan olivat T4-ympäristön yhdistelmäkyselyssä JA-operaattoria käytettäessä reilut 2 prosenttiyksikköä ja virkeoperaattoria käytettäessä noin yhden prosenttiyksikön verran korkeammat kuin perinteisellä tavalla haettaessa. Virkeoperaattorin tapauksessa ero ei ollut tilastollisesti merkitsevä eikä varsinaisesti JA-operaattorillakaan (ero oli merkitsevä vasta merkitsevyystasolla 0.1 eli käytännössä olematon). Tämä ei vastaa hypoteesia 6, jonka mukaan T4-ympäristön tarkkuusarvojen pitäisi olla paremmat kuin perinteisen yhdistelmäkyselyn tarkkuusarvojen.

Kun T4-ympäristön yhdistelmäkyselyn tarkkuusarvoja verrattiin saman T4-ympäristön johdoskyselyn (AB) ja yhdyssanakyselyn (AC) tarkkuusarvoihin, erot eivät virkeoperaattorin tapauksessa olleet tilastollisesti merkitseviä eivätkä käytännössä JA-operaattorinkaan tapauksessa (vain lievästi: yhdyssanakyselyn ja yhdistelmäkyselyn välillä vasta merkitsevyystasolla 0.1).

T4-tutkimusympäristön perusjoukon yhdistelmäkyselyssä siis kävi niin, että haun tarkkuus nousi hiukan perinteistä hakutapaa korkeammaksi samalla kun haun saanti oli sama kuin perinteisellä tavalla haettaessa. Ero ei kuitenkaan ollut tilastollisesti merkitsevä. Tällainen lopputulos kuitenkin edellyttää sitä, että perusmuotohakemistosta muistetaan hakea myös hakusanan johdosperhe ja kaikki hakusanalla alkavat yhdyssanat.

Kaiken kaikkiaan T4-ympäristössä vain peruskyselyn (A) tarkkuusarvot olivat selvästi paremmat kuin T1-ympäristön yhdistelmäkyselyn. Jos haun tarkkuus on ehdottoman tärkeää, hakijan siis kannattaisi käyttää kyselyissä vain hakusanan perusmuotoja. – Jos tarkkuudesta voidaan tinkiä, kannattaa saman tien hakea perusmuotojen lisäksi myös johdoksilla, yhdyssanoilla ja mieluiten molemmilla, koska siten saanti paranee selvästi ilman, että tarkkuus erityisemmin kärsii.

9.4.2 Johdososajoukko

Peruskysely

Kun tutkimusympäristössä T4 tarkasteltiin vain johdososajoukon hakuja (taulukko 24; liite 14, luku 2), peruskyselyn (A) suhteellinen **saanti** oli JA-operaattoria käytettäessä 31 ja virkeoperaattoria käytettäessä 22 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn (ABC). Molempien operaattorien tapauksessa ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01. Erot olivat käytännössä olennaiset. Tämä oli täysin hypoteesin 4 mukaista.

Kun peruskyselyä verrattiin saman T4-tutkimusympäristön muiden kyselytyyppien kanssa, **saantiarvojen** väliset erot olivat sekä JA- että virkeoperaattoria käytettäessä tilastollisesti merkitseviä merkitsevyystasolla 0.01 myös silloin, kun vertailukohteena oli johdoskysely (AB) tai yhdistelmäkysely (ABC). Erot olivat näissä tapauksissa käytännössä olennaiset.

Peruskyselyn **tarkkuus** puolestaan oli 20 prosenttiyksikköä (JA-operaattori) tai 24 prosenttiyksikköä (virkeoperaattori) korkeampi kuin perinteisen

yhdistelmäkyselyn (ABC) tarkkuus. Tämä ero oli JA-operaattorin tapauksessa merkitsevä merkitsevyystasolla 0.01 ja virkeoperaattoria käytettäessä merkitsevyystasolla 0.05. Molempien operaattorien tapauksessa erot olivat käytännössä olennaiset (Sparck Jonesin mittari). Tämä oli täysin hypoteesin 4 mukaista.

Kun peruskyselyä verrattiin saman T4-tutkimusympäristön muiden kyselytyyppien kanssa, sen ja yhdistelmäkyselyn (ABC) tarkkuusarvojen väliset erot olivat JA-operaattoria käytettäessä merkitseviä merkitsevyystasolla 0.01. Virkeoperaattoria käytettäessä tarkkuusarvojen väliset erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.05, kun peruskyselyn vertailukohteina olivat oli yhdyssanakysely (AC) tai yhdistelmäkysely (ABC).

Johdososajoukossa siis peruskyselyn tulosjoukon tarkkuusarvo voittaa selkeästi vain yhdistelmäkyselyn tulosjoukon tarkkuusarvon. Sen paremmuus johdos- tai yhdyssanakyselyihin verrattuna ei ole läheskään niin selvä. Sen sijaan peruskyselyn tulosjoukkojen saantiarvo oli selvästi huonompi kuin muiden kyselytyyppien. Niinpä tämän perusteella voidaan sanoa, että johdososajoukossa ei kannata hakea pelkästään hakusanan perusmuodoilla - jos hakusanaan liittyy johdoksia, ne kannattaa lisätä kyselyyn.

Taulukko 24. T1- ja T4-ympäristöissä kyselytyypeillä AC ja ABC saatujen hakutulosten väliset erot johdososajoukossa (N = 8).

Operaattori	Kyselytyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T4**	**T1**	**T4**	**T1**	**T4**
JA	A		12,3		51,9		67,3
	AB		27,1		74,5		56,5
	AC		17,6		57,7		57,2
	ABC	**41,3**	**41,3**	**82,7**	**83,4**	**46,9**	**48,2**
Virke	A		4,4		34,3		88,6
	AB		11,3		52,8		76,4
	AC		6,3		38,0		72,7
	ABC	**15,9**	**15,9**	**56,2**	**56,9**	**64,3**	**66,8**

A	Peruskysely
AB	Johdoskysely
AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat

Johdoskysely

T4-ympäristön johdoskyselyn (AB) **saanti** jäi JA-operaattoria käytettäessä 8 ja virkeoperaattoria käytettäessä reilut 3 prosenttiyksikköä alemmaksi kuin perinteisen yhdistelmäkyselyn (ABC). Edellämainitut saantiarvojen väliset erot olivat tilastollisesti merkitseviä: JA-operaattorin tapauksessa merkitsevyystasolla 0.01 ja virkeoperaattoria käytettäessä merkitsevyystasolla 0.05. Ero oli JA-operaattorin osalta myös käytännössä huomattava. Tulos ei siis tue hypoteesia 5, koska sen mukaan saannin pitäisi olla keskimäärin sama.

T4-ympäristön johdoskyselyn **tarkkuus** oli perinteistä yhdistelmäkyselyä parempi: eroa oli JA-operaattoria käytettäessä vajaat 10 ja virkeoperaattoria käytettäessä 12 prosenttiyksikköä. Nämä tarkkuusarvojen väliset erot olivat tilastollisesti merkitseviä sekä JA-operaattorilla että virkeoperaattorilla merkitsevyystasolla 0.05. JA-operaattorilla tulos oli käytännössä huomattava ja virkeoperaattorilla olennainen. Tämä tulos oli hypoteesin 5 mukainen eli johdoskyselyn tarkkuuden pitäisikin olla parempi kuin perinteisen yhdistelmäkyselyn.

Johdoskyselyn saantiarvot siis jäävät peruskyselyn ja yhdistelmäkyselyn väliin; kumpaankin suuntaan verrattuna saantiarvojen välinen ero on tilastollisesti merkitsevä. Johdoskyselyn tarkkuusarvo puolestaan on selvästi yhdistelmäkyselyn tarkkuusarvoa parempi, mutta ei juuri peruskyselyn tarkkuusarvoa huonompi. Näin ollen peruskysely kannattaisi aina laajentaa johdosperheellä, koska silloin saanti paranee useita prosenttiyksikköjä enemmän kuin tarkkuus huononee.

Yhdyssanakysely

T4-ympäristön yhdyssanakyselyn (AC) **saantiarvo** oli JA-operaattorilla 25 ja virkeoperaattorilla 18 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn (ABC). Eri tutkimusympäristöjen saantiarvojen väliset erot olivat kummankin operaattorin tapauksessa tilastollisesti merkitseviä merkitsevyystasolla 0.01. Käytännön tasolla erot olivat olennaiset (Sparck Jonesin mittari). Pelkkä hakusanojen automaattinen katkaisu ja peruskyselyn laajentaminen näin saaduilla vartaloilla yhdyssanakyselyksi ei siis johdososajoukossa riitä nostamaan saantiarvoa perinteisen yhdistelmäkyselyn saantiarvojen tasolle.

Yhdyssanakyselyn (AC) **tarkkuus** oli JA-operaattoria käytettäessä 10 ja virkeoperaattoria käytettäessä 8 prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn (ABC). JA-operaattorin tapauksessa tarkkuusarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01 ja Sparck Jonesin mittarilla käytännössä olennainen. Virkeoperaattoria käytettäessä ero ei ollut tilastollisesti merkitsevä.

Johdososajoukossa yhdyssanakyselyn saanti siis on vain hieman parempi kuin peruskyselyn, kun taas tarkkuus on selvästi peruskyselyä huonompi. Kokonaisuuden kannalta on siis suositeltavaa laajentaa yhdyssanakysely yhdistelmäkyselyksi.

Yhdistelmäkysely

Kun T4-ympäristön yhdistelmäkyselyä (ABC) verrattiin perinteiseen yhdistelmäkyselyyn, sen tulosjoukoissa sekä **saanti-** että **tarkkuusarvot** olivat hivenen perinteisen yhdistelmäkyselyn vastaavia arvoja paremmat. Tutkimusympäristöjen väliset erot eivät kuitenkaan olleet tilastollisesti merkitseviä. T4-ympäristössä siis päästään haluttaessa yhtä hyviin - tai itse asiassa hiukan parempiinkin - tuloksiin kuin T1-ympäristössä. Vaikka tulos sinänsä onkin hypoteesin 6 mukainen, sitä ei kuitenkaan voi vahvistaa tilastollisten merkitsevyystestien avulla eli erot eivät olleet riittävän selvät. Näin ollen hypoteesille 6 ei saatu riittävästi tukea.

Johdososajoukossa saannin ja tarkkuuden käänteisyys näkyy selkeämmin kuin perusjoukossa. JA-operaattoria käytettäessä käy niin, että kun siirrytään suppeammasta kyselytyypistä laajempaan, saanti kasvaa useita prosenttiyksikköjä enemmän kuin tarkkuus laskee: kun kysely laajennetaan peruskyselystä (A) yhdistelmäkyselyyn (ABC), saanti kasvaa kokonaista 32 prosenttiyksikköä, kun tarkkuus samalla alenee 19 prosenttiyksikköä. Kyselyn laajentaminen siis on kannattavaa, koska saanti paranee enemmän kuin tarkkuus kärsii.

Vaikka saantia johdososajoukossa kasvatetaan parhaiten laajentamalla peruskysely johdoskyselyksi (AB), on suurimman saantiarvon saamiseksi laajennettava kyselyä myös yhdyssanoilla. Hypoteesi 5 ei siis saanut riittävästi tukea, ei siis myöskään hypoteesi 6.

9.4.3 Yhdyssanaosajoukko

Osittamattomien ja ositettujen hakusanojen väliset erot

Taulukossa 25 näkyvät vertailut yhdyssanaosajoukossa eli siinä yhdeksän kyselyn osajoukossa, jossa hakusanat olivat yhdyssanoja. T4-ympäristössä yhdyssanojen osilla hakeminen tarkoittaa katkaistujen hakusanojen käyttöä, jolloin löydetään yhdyssanojen alkuosina esiintyneet hakusanat – yhdyssanojen keski- ja loppuosiin ei T4-ympäristössä päästy käsiksi.

Kaikki ne kyselytyypit, joissa yhdyssanoja ei jaettu osiinsa (T1/ABC, T4/A, T4/AB, T4/AC ja T4/ABC) olivat **saanniltaan** selvästi huonompia kuin osien yhdistelmäkyselyt (T1/ABCabc, T4/ABCabc), osien yhdyssanakysely (T4/ACac) ja osien johdoskysely (ABab).

Edellämainittujen ryhmien saantiarvojen väliset erot olivat niin JA- kuin virkeoperaattoriakin käytettäessä tilastollisesti merkitseviä, enimmäkseen merkitsevyystasolla 0.01. Heikommalla merkitsevyystasolla 0.05 erosi osien johdoskysely (ABab) yhdistelmäkyselyistä (T1/ABC ja T4/ABC). Tämä päti sekä JA- että virkeoperaattoria käytettäessä. Lisäksi heikommalla merkitsevyystasolla eli tasolla 0.05 erosivat virkeoperaattoria käytettäessä johdoskysely (AB) verrattuna osien johdoskyselyyn (ABab) sekä toisaalta yhdistelmäkysely (ABC) verrattuna osien yhdyssanakyselyyn (ACac). (Liite 14, luku 3.)

Osien peruskysely (Aa) ei pärjännyt vertailuissa yhtä hyvin kuin toiset yhdyssanojen osia sisältävät kyselytyypit: vaikka sen saantiarvot olivatkin paremmat kuin minkään osittamattomia yhdyssanoja sisältävien kyselytyyppien, saantiarvojen väliset erot olivat vain muutamassa tapauksessa tilastollisesti merkitseviä: JA-operaattoria käytettäessä osien peruskysely erosi peruskyselystä (A) merkitsevyystasolla 0.01 ja toisaalta johdoskyselystä (AB) merkitsevyystasolla 0.05. Virkeoperaattoria käytettäessä osien peruskysely erosi vain peruskyselystä, silloin merkitsevyystasolla 0.05.

Kun osittamattomia yhdyssanoja sisältävien kyselytyyppien saantiarvoja verrataan keskenään, näiden kesken ei virkeoperaattoria käytettäessä ole tilastollisesti merkitseviä eroja, mutta JA-operaattoria käytettäessä peruskysely (A) erosi yhdistelmäkyselyistä (T1/ABC, T4/ABC) merkitsevyystasolla 0.05. Ositettuja hakusanoja sisältävien kyselytyyppien keskinäisessä

Taulukko 25. T1- ja T4-ympäristöissä kaikilla kyselytyypeillä saatujen hakutulosten väliset erot yhdyssanaosajoukossa (N = 9).

Ope-raattori	Kysely-tyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T4**	**T1**	**T4**	**T1**	**T4**
JA	A		6,1		48,8		86,1
	AB		6,7		51,2		84,7
	AC		7,0		54,5		80,7
	ABC	8,0	7,4	55,7	56,3	76,1	79,4
	Aa		11,9		67,4		65,7
	ABab		14,3		73,0		56,1
	ACac		21,8		83,7		49,5
	ABCabc	**25,8**	**24,9**	**88,1**	**88,6**	**42,1**	**44,6**
Virke	A		5,6		47,1		87,5
	AB		5,9		48,8		87,5
	AC		5,8		47,7		82,3
	ABC	6,1	6,1	48,8	49,4	80,0	82,3
	Aa		6,8		56,1		89,9
	ABab		7,2		58,5		90,2
	ACac		8,1		61,6		83,7
	ABCabc	**8,6**	**8,4**	**62,8**	**63,4**	**81,1**	**83,7**

A	Peruskysely
AB	Johdoskysely
AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
Aa	Osien peruskysely
ABab	Osien johdoskysely
ACac	Osien yhdyssanakysely
ABCabc	Osien yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat

vertailussa löytyi tilastollisesti merkitseviä eroja, joista tarkemmin seuraavissa alaluvuissa.

Saannin ja tarkkuuden käänteisyys näkyy eri kyselytyyppien **tarkkuusarvoja** vertailtaessa: saanniltaan parhaat kyselytyypit eli T1- ja T4-ympäristöjen osien yhdistelmäkyselyt (ABCabc) ja osien yhdyssanakysely (T4/ACac) olivat JA-operaattoria käytettäessä tarkkuudeltaan huonoimmat – erot osittamattomiin kyselytyyppeihin verrattuna olivat tilastollisesti merkitsevät merkitsevyystasolla 0.01 (poikkeuksena vain T4-ympäristön osien yhdyssanakyselyn ja T1-ympäristön yhdistelmäkyselyn ABC saantiarvojen ero,

joka ei ollut tilastollisesti merkitsevä). Osien johdoskyselyn (ABab) erot osittamattomiin kyselytyyppeihin taas olivat useimmissa tapauksissa merkitseviä merkitsevyystasolla 0.05.

Toisaalta virkeoperaattoria käytettäessä eri kyselytyyppien välillä ei ollut tilastollisesti merkitseviä eroja. Virkeoperaattoria käytettäessä siis kannattaa jakaa yhdyssanat osiinsa, koska saanti paranee ilman että tarkkuus erityisemmin kärsii.

Tulosten perusteella yhdyssanojen osittaminen ja osilla hakeminen on syytä tehdä aina, kun halutaan saada mahdollisimman hyvä saanti, koska osittaminen selvästi parantaa saantiarvoja. Saannin paranemiselle löytyy enemmän tilastollista tukea (korkeampi merkitsevyystaso) kuin tarkkuuden heikkenemiselle.

Osien peruskysely

Osien peruskyselyn (Aa) **saanti** oli JA-operaattoria käytettäessä 20 prosenttiyksikköä ja virkeoperaattoria käytettäessä 7 prosenttiyksikköä alempi kuin perinteisen osien yhdistelmäkyselyn (ABCabc). JA-operaattorin tapauksessa saantiarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01, virkeoperaattorin tapauksessa merkitsevyystasolla 0.05. Ero oli käytännössä JA-operaattorin tapauksessa olennainen, virkeoperaattorilla huomattava. Tämä tulos oli hypoteesin 4 mukainen.

Kun osien peruskyselyn saantia verrattiin muiden saman T4-ympäristön kyselytyyppien saantiarvoihin, tilastollisesti merkitseviä eroja löytyi, kun vertailukohteina olivat osien yhdyssanakysely (ACac) ja osien yhdistelmäkysely (ABCabc). JA-operaattoria käytettäessä saannin erot olivat molemmissa tapauksissa merkitseviä merkitsevyystasolla 0.01. Virkeoperaattoria käytettäessä osien peruskysely erosi osien yhdyssanakyselystä merkitsevyystasolla 0.05 ja osien yhdistelmäkyselystä merkitsevyystasolla 0.01.

T4-ympäristön osien peruskyselyn (Aa) **tarkkuus** oli 24 (JA-operaattori) tai 9 prosenttiyksikköä (virkeoperaattori) parempi kuin perinteisen osien yhdistelmäkyselyn (ABCabc). JA-operaattoria käytettäessä tarkkuusarvojen välinen ero oli merkitsevä merkitsevyystasolla 0.01, sen sijaan virkeoperaattorilla ero ei ollut tilastollisesti merkitsevä. JA-operaattorin tapauksessa ero oli käytännössä olennainen. JA-operaattorin osalta tulos oli hypoteesin 4 mukainen, virkeoperaattorilla ei.

Kun osien peruskyselyn tarkkuusarvoja verrattiin muiden saman T4-ympäristön kyselytyyppien kanssa, eri kyselytyyppien tarkkuusarvojen väliset erot eivät virkeoperaattoria käytettäessä olleet tilastollisesti merkitseviä. Sen sijaan JA-operaattoria käytettäessä osien peruskyselyn tarkkuus erosi tilastollisesti merkitsevästi merkitsevyystasolla 0.01 saman T4-ympäristön osien yhdistelmäkyselyn (ABCabc) tarkkuudesta.

Osien yhdyssanakysely

Osien yhdyssanakyselyn (ACac) **saanti** jäi JA-operaattorilla 4 ja virkeoperaattorilla vain prosenttiyksikön verran alemmaksi kuin perinteisen osien yhdistelmäkyselyn (ABCabc). Edellämainitut saantiarvojen väliset erot eivät olleet tilastollisesti merkitseviä.

Osien yhdyssanakyselyn **tarkkuus** T4-ympäristössä oli JA-operaattorilla reilut 7 ja virkeoperaattorilla vajaat 3 prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn tarkkuus. Tarkkuusarvojen väliset erot olivat JA-operaattoria käytettäessä tilastollisesti merkitseviä merkitsevyystasolla 0.05. Tässä tapauksessa ero oli myös käytännössä huomattava.

Yhdyssanat siis kannattaa jakaa osiinsa ja osat lisätä kyselyyn alkuperäisen hakusanan rinnalle. Vaikka saanti jää alle perinteisen osien yhdistelmäkyselyn saannin, ero ei ole tilastollisesti merkitsevä; sen sijaan tarkkuus on selvästi parempi kuin perinteisen osien yhdistelmäkyselyn, ja tämä ero on myös tilastollisesti merkitsevä. Toisaalta osien yhdyssanakyselyn saanti on selvästi korkeampi kuin osien peruskyselyn (ero tilastollisesti merkitsevä), mutta tarkkuus ei jää olennaisesti jälkeen osien peruskyselyn tarkkuudesta (ero ei tilastollisesti merkitsevä). T4-ympäristössä ei siis kannata käyttää osien peruskyselyä, vaan laajentaa se saman tien osien yhdyssanakyselyksi.

Osien johdoskysely

Osien johdoskyselyn (ABab) **saanti** jäi JA-operaattoria käytettäessä 15 ja virkeoperaattoria käytettäessä 4 prosenttiyksikköä alemmaksi kuin perinteisen osien yhdistelmäkyselyn (ABCabc). Saantiarvojen välinen ero oli JA-operaattoria käytettäessä tilastollisesti merkitsevä merkitsevyystasolla 0.01 (ja käytännössä olennainen). Virkeoperaattoria käytettäessä saantiarvojen välinen ero ei ollut tilastollisesti merkitsevä. Virkeoperaattorin tapauksessa tulos oli hypoteesin 5 mukainen, JA-operaattorilla taas ei.

Kun osien johdoskyselyn saantiarvoja verrattiin saman T4-ympäristön muiden kyselytyyppien kanssa, saanti erosi JA-operaattoria käytettäessä tilastollisesti merkitsevästi osien yhdistelmäkyselystä (ABCabc; merkitsevyystasolla 0.01) ja osien yhdyssanakyselystä (ACac; merkitsevyystasolla 0.05). Virkeoperaattoria käytettäessä saantiarvot eivät eronneet toisistaan niin paljon, että tilastollisesti merkitseviä eroja olisi löytynyt.

Osien johdoskyselyn **tarkkuusarvot** olivat JA-operaattorilla 14 ja virkeoperaattorilla 9 prosenttiyksikköä korkeammat kuin perinteisen osien yhdistelmäkyselyn. JA-operaattorin tapauksessa tarkkuusarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01; sen sijaan virkeoperaattorilla ero ei ollut tilastollisesti merkitsevä. Sparck Jonesin määrittelyjen perusteella ero oli JA-operaattoria käytettäessä olennainen. JA-operaattorin osalta tulos oli hypoteesin 5 mukainen, virkeoperaattorilla ei, eli virkeoperaattori rajaa hakutulosta jo muutenkin tarkaksi.

Kun osien johdoskyselyn tarkkuusarvoja verrattiin muiden saman T4-ympäristön kyselytyyppien kanssa, olivat tarkkuusarvojen väliset erot tilastollisesti merkitseviä merkitsevyystasolla 0.05, kun vertailukohteena oli osien yhdistelmäkysely (ABCabc) ja hakusanat oli yhdistetty toisiinsa JA-operaattorilla. Virkeoperaattoria käytettäessä tarkkuusarvojen väliset erot eivät olleet tilastollisesti merkitseviä.

Kun osien peruskysely siis laajennettiin osien johdoskyselyksi, JA-operaattoria käytettäessä saanti parani jonkin verran, mutta tarkkuus heikkeni useampia prosenttiyksikköjä enemmän kuin saanti kasvoi. Virkeoperaattoria käytettäessä näiden kahden kyselytyypin saanti- ja tarkkuusarvojen välillä ei ollut suurtakaan eroa. Yhdyssanaosajoukossa johdosten lisäämisellä ei siis ole niin suurta merkitystä kuin perusjoukossa ja johdososajoukossa.

Osien yhdistelmäkysely

Osien yhdistelmäkyselyn (ABCabc) **saanti** nousi sekä JA- että virkeoperaattorilla hienokseltaan (alle prosenttiyksikön verran) yli perinteisen yhdistelmäkyselyn saannin. Saantiarvojen väliset erot eivät olleet tilastollisesti merkitseviä. Osien yhdistelmäkyselyn (ABCabc) **tarkkuus** oli T4-ympäristössä molemmilla operaattoreilla vajaat 3 prosenttiyksikköä korkeampi kuin perinteisen osien yhdistelmäkyselyn tarkkuus. Näiden kahden tutkimusympäristön tarkkuusarvojen välinen ero ei ollut tilastollisesti merkitsevä. Vaik-

ka tulos siis olikin hypoteesin 6 mukainen, sille ei saatu kuitenkaan vahvistusta, koska erot eivät olleet tilastollisesti merkitseviä.

Yhteenveto

Osien peruskyselyn laajentaminen yhdyssanojen osilla osien yhdyssanakyselyksi riittää nostamaan T4-ympäristön tulosjoukkojen saantiarvot lähes T1-ympäristön osien yhdistelmäkyselyn tasalle. Osien yhdistelmäkyselyn tasalle pääsemiseksi - itse asiassa sen ylittämiseksi - on kyselyä kuitenkin laajennettava sekä yhdyssanoilla että johdosperheellä.

Saantiarvojen parantamiseksi kannattaisi osien peruskyselyjä laajentaa automaattisesti hakusanan sisältävillä yhdyssanoilla, koska ero pelkän osien peruskyselyn saantiarvoihin on tuolloin huomattava ja lähes yhtä hyvä kuin osien yhdistelmäkyselyn tulosjoukon saantiarvo. Samalla osien yhdyssanakyselyn tulosjoukon tarkkuusarvo on - erityisesti JA-operaattorin tapauksessa - selvästi parempi kuin osien yhdistelmähaun.

Siis: jos yhdyssanan jakaa osiin, ei kannata hakea pelkillä osien perusmuodoilla, vaan hakea myös dokumentit, joihin sisältyy hakusanan osia sisältäviä sanaliittoja ja yhdyssanoja.

T4-ympäristössä kyselyn laajentaminen johdosperheen perusmuodoilla ei huononna tarkkuutta niin paljon kuin vastaava laajennus sellaisissa tutkimusympäristöissä, joissa hakusanat katkaistaan (esimerkiksi T2-ympäristössä automaattisesti vartalo-ohjelmien avulla). Tämän perusteella kyselyn laajentaminen hakutesauruksen avulla ilmeisesti toimii perusmuotohakemistossa paremmin (tarkemmin) kuin taivutusmuotohakemistossa. Perusmuotohakemistossa voidaan käyttää sanojen täsmällisiä muotoja, jotka eivät tuota tulosjoukkoon niin paljon epärelevantteja dokumentteja kuin katkaistut hakusanat.

Kyselyjen laajentaminen T4-ympäristössä edellä kuvatuin menetelmin yleensä nostaa saantiarvoja ja laskee tarkkuusarvoja jokseenkin yhtä monta prosenttiyksikköä. Käytännössä tämä merkitsee sitä, että käyttäjä saa - tai hänen on pakko - valita, haluaako hän painottaa haun saantia vai tarkkuutta ja sitten laajennettava tai supistettava kyselyä valintansa mukaisesti (esimerkiksi lisäämällä tai poistamalla kyselystä yhdyssanoja tai hakusanojen johdoksia).

9.5 Taivutusmuotohakemistosta ja ositetusta perusmuotohakemistosta saatujen hakutulosten vertailu

Tutkimusympäristössä T5 eli ositetussa perusmuotohakemistossa peruskysely (A) ja johdoskysely (AB) toteutettiin samalla tavalla kuin edellisessä luvussa kuvatussa T4-tutkimusympäristössä. Yhdyssanojen osat haettiin eri tavalla eli niitä ei katkaistu vartalo-ohjelmien avulla kuten T4-ympäristössä, vaan yhdyssanojen eri osiin liitettiin katkaisumerkki. Lisäksi T5-ympäristössä voitiin hakea myös yhdyssanojen keski- ja loppuosia.

9.5.1 Perusjoukko

Peruskysely

T5-ympäristössä peruskyselyn (A) tulosjoukon suhteellinen **saanti** oli JA-operaattoria käytettäessä 17 ja virkeoperaattoria käytettäessä 14 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn (ABC) tulosjoukon (taulukko 26). Kyselytyyppien välinen ero oli sekä JA- että virkeoperaattoria käytettäessä tilastollisesti merkitsevä merkitsevyystasolla 0.01. Sparck Jonesin määrittelyn mukaan erot olivat käytännössä olennaisia. (Liite 14, luku 1.) Tämä tulos on linjassa hypoteesin 4 kanssa.

Peruskyselyn tulosjoukon saantiarvot olivat myös selvästi alemmat kuin muiden T5-ympäristön kyselytyyppien tulosjoukkojen saantiarvot. JA-operaattorin tapauksessa erot verrattuna yhdyssanakyselyyn (AC) ja yhdistelmäkyselyyn (ABC) olivat tilastollisesti merkitseviä merkitsevyystasolla 0.01 (käytännössä olennaiset) ja verrattuna johdoskyselyyn (AB) merkitsevyystasolla 0.05 (käytännössä huomattava).

Virkeoperaattoria käytettäessä erot peruskyselyn ja muiden T5-ympäristön kyselytyyppien saantiarvojen välillä olivat tilastollisesti merkitseviä merkitsevyystasolla 0.01. Erot peruskyselyn ja johdoskyselyn sekä peruskyselyn ja yhdyssanakyselyn välillä olivat käytännössä huomattavat; ero peruskyselyn ja yhdistelmäkyselyn saantiarvojen välillä oli käytännössä olennainen.

Peruskyselyn (A) tulosjoukon **tarkkuus** oli sekä JA- että virkeoperaattoria käytettäessä 8 prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn (ABC) tulosjoukon. JA-operaattorin tapauksessa tarkkuusarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01; lisäksi ero oli käytännössä huomattava. Virkeoperaattoria käytettäessä ei tarkkuus-

Taulukko 26. T1- ja T5-ympäristöissä kyselytyypeillä A, AB, AC ja ABC saatujen hakutulosten väliset erot perusjoukossa (N = 26).

Ope-raattori	Kysely-tyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T5**	**T1**	**T5**	**T1**	**T5**
JA	A		8,9		54,7		75,8
	AB		13,5		61,7		72,4
	AC		12,7		65,7		73,2
	ABC	**20,0**	**19,8**	**72,0**	**75,1**	**68,0**	**71,0**
Virke	A		6,0		46,3		84,2
	AB		8,1		52,0		80,4
	AC		8,3		54,6		79,2
	ABC	**11,3**	**11,3**	**60,1**	**61,0**	**76,6**	**77,3**

A	Peruskysely
AB	Johdoskysely
AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat

arvojen välillä ollut tilastollisesti merkitseviä eroja. JA-operaattorin osalta tulos on linjassa hypoteesin 4 kanssa.

Myös peruskyselyn ja T5-ympäristön yhdistelmäkyselyn tarkkuusarvojen välinen ero oli JA-operaattorin tapauksessa tilastollisesti merkitsevä merkitsevyystasolla 0.05. Ero ei kuitenkaan ollut käytännössä merkittävä. Muuten peruskyselyn ja T5-ympäristön muiden kyselytyyppien tarkkuusarvojen välille ei syntynyt tilastollisesti merkitseviä eroja.

Siis: peruskyselyn tulosjoukkojen saantiarvot jäivät selvästi huonommiksi kuin muiden kyselytyyppien saantiarvot, mutta niiden tarkkuusarvot eivät kuitenkaan vastaavassa määrin nousseet toisten kyselytyyppien tarkkuusarvoja paremmiksi. Peruskyselyä ei ositetussa perusmuotohakemistossa siis käytännössä kannata käyttää.

Johdoskysely

Johdoskyselyn (AB) **saanti** oli JA-operaattorilla 10 ja virkeoperaattorilla 8 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn (ABC). Sekä JA- että virkeoperaattorin tapauksessa saantiarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01. Sparck Jonesin määrittelyn mu-

kaan ero oli JA-operaattorilla käytännössä olennainen, virkeoperaattorilla huomattava. Tämä ei ole linjassa hypoteesin 5 kanssa; sen mukaan saantiarvojen piti olla keskimäärin samat.

Kun johdoskyselyä verrattiin muihin T5-ympäristön kyselytyyppeihin, saantiarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01, kun vertailukohteena oli yhdistelmäkysely (ABC). Tämä tulos saatiin sekä JA- että virkeoperaattoria käytettäessä. Ero oli JA-operaattorin tapauksessa käytännössä olennainen ja virkeoperaattorilla käytännössä huomattava.

Johdoskyselyn **tarkkuus** oli sekä JA- että virkeoperaattorilla 4 prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn tarkkuus. Tarkkuusarvojen välinen ero oli JA-operaattorin tapauksessa tilastollisesti merkitsevä merkitsevyystasolla 0.01, sen sijaan virkeoperaattoria käytettäessä ero ei ollut tilastollisesti merkitsevä. JA-operaattorin tapauksessa tulos oli hypoteesin 5 mukainen - tosin Sparck Jonesin määrittelyn perusteella erolla ei ole käytännön merkitystä.

Perusjoukossa peruskyselyn lisäksi siis myös johdoskyselyn saanti jäi melko alhaiseksi, vaikka laajennus johdosperheen jäsenillä paransikin saantia verrattuna pelkkään peruskyselyyn. Johdosten lisääminen kuitenkin paransi saantiarvoja useita prosenttiyksikköjä enemmän kuin samalla alensi tarkkuusarvoja.

Yhdyssanakysely

Yhdyssanakyselyn (AC) suhteellinen **saanti** jäi sekä JA- että virkeoperaattoria käytettäessä suunnilleen 6 prosenttiyksikköä alemmaksi kuin perinteisen yhdistelmäkyselyn (ABC) saanti. Saantiarvojen välinen erot olivat tilastollisesti merkitseviä: JA-operaattorilla merkitsevyystasolla 0.05 ja virkeoperaattorilla merkitsevyystasolla 0.01. Erot olivat käytännössä huomattavat.

Kun yhdyssanakyselyn saantiarvoja verrattiin saman tutkimusympäristön muiden kyselytyyppien saantiarvoihin, tilastollisesti merkitseviä eroja todettiin, kun vertailukohteena oli yhdistelmäkysely (ABC) tai peruskysely (A). Erot olivat sekä JA- että virkeoperaattoria käytettäessä merkitseviä merkitsevyystasolla 0.01. Käytännössä - Sparck Jonesin mittarilla - erot olivat enimmäkseen huomattavat; peruskyselyn ja yhdyssanakyselyn välinen ero oli kuitenkin JA-operaattoria käytettäessä käytännössä olennainen. Yhdys-

sanakyselyn saanti siis oli selvästi peruskyselyn saantia parempi, mutta toisaalta selvästi yhdistelmäkyselyä huonompi.

Yhdyssanakyselyn **tarkkuus** puolestaan oli JA-operaattoria käytettäessä 5 ja virkeoperaattoria käytettäessä 3 prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn. Tarkkuusarvojen välinen ero oli JA-operaattoria käytettäessä tilastollisesti merkitsevä merkitsevyystasolla 0.01 ja Sparck Jonesin mittarin perusteella nipin napin huomattava. Virkeoperaattoria käytettäessä tarkkuusarvojen väliset erot eivät olleet tilastollisesti merkitseviä.

Koska peruskysely voidaan helposti laajentaa automaattisesti yhdyssanakyselyksi eli lisäämällä kyselyyn hakusanan rinnalle yhdyssanan osat, tämä kannattaa T5-ympäristön perusjoukossa tehdä. Vaikka tarkkuus alenee joitakin prosenttiyksikköjä, saanti samalla nousee selvästi useamman prosenttiyksikön verran. Tämä pätee erityisesti silloin, kun hakusanat on kytketty toisiinsa virkeoperaattorilla, jolloin tulosjoukkojen tarkkuusarvojen välillä ei ole sanottavia eroja.

Yhdistelmäkysely

Kun yhdyssanakyselyä edelleen laajennettiin johdoksilla eli toteutettiin yhdistelmäkysely (ABC), T5-ympäristö osoittautui T1-ympäristöä jonkin verran paremmaksi sekä saannin että tarkkuuden suhteen: JA-operaattoria käytettäessä sekä **tarkkuus** että suhteellinen **saanti** olivat T5-ympäristössä 3 prosenttiyksikköä korkeammat kuin T1-ympäristössä. Virkeoperaattoria käytettäessä taas T5-ympäristön **saanti**- ja **tarkkuusarvot** olivat vajaan prosenttiyksikön verran paremmat kuin T1-tutkimusympäristössä. Mitkään edellämainituista eroista eivät olleet tilastollisesti merkitseviä.

Hypoteesissa 7 väitettiin T5-ympäristön saannin olevan paremman ja tarkkuuden huonomman kuin T1-ympäristön. Vaikka näin saannin osalta kävikin ja toisaalta taas T5-ympäristön tarkkuus nousikin keskimäärin paremmaksi kuin T1-ympäristön tarkkuus, erot eivät olleet niin suuret, että ne olisivat olleet tilastollisesti merkitseviä. Nyt JA-operaattoria käytettäessä tarkkuusarvojen ero oli tilastollisesti merkitsevä vain merkitsevyystasolla 0.1, mikä ei käytännössä riitä.

T5-tutkimusympäristön perusjoukon yhdistelmäkyselyssä siis tarkkuus ja saanti nousivat hiukan T1-ympäristön vastaavia arvoja korkeammiksi. Tällainen lopputulos kuitenkin edellyttää sitä, että perusmuotohakemistosta

muistetaan hakea myös hakusanan johdosperhe ja kaikki hakusanan sisältävät yhdyssanat.

9.5.2 Johdososajoukko

Peruskysely

Kun tarkasteltiin tarkemmin sitä kahdeksan kyselyn osajoukkoa, jossa kyselyjä voitiin laajentaa johdosperheen jäsenillä (taulukko 27; liite 14, luku 2), peruskyselyn (A) suhteellinen **saanti** oli JA-operaattoria käytettäessä 31 ja virkeoperaattoria käytettäessä 22 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn (ABC). Molempien operaattoreiden tapauksessa saantiarvojen välinen ero oli merkitsevä merkitsevyystasolla 0.01. Erot olivat käytännössä olennaiset. Tulos vastaa hypoteesin 4 väitettä.

Myös silloin, kun peruskyselyä verrattiin saman T5-ympäristön muiden kyselytyyppien kanssa, olivat saantiarvojen väliset erot JA-operaattorin tapauksessa tilastollisesti merkitseviä merkitsevyystasolla 0.01, kun vertailukohteena olivat johdoskysely (AB) ja yhdistelmäkysely (ABC) ja merkitsevyystasolla 0.05, kun sitä verrattiin yhdyssanakyselyyn (AC). Kahdessa ensimmäisessä erot olivat käytännössä olennaiset, yhdyssanakyselyyn verrattaessa käytännössä huomattava.

Virkeoperaattoria käytettäessä saantiarvojen välille löytyi tilastollisesti merkitseviä eroja merkitsevyystasolla 0.01, kun peruskyselyn vertailukohteina olivat johdoskysely (AB) ja yhdistelmäkysely (ABC). Kummassakin tapauksessa ero oli Sparck Jonesin mittarilla käytännössä olennainen. Sekä JA- että virkeoperaattorin tapauksessa peruskyselyn saanti siis jäi varsin kehnoksi.

Peruskyselyn **tarkkuus** oli JA-operaattorilla 20 ja virkeoperaattorilla 24 prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn (ABC). JA-operaattorilla ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01, virkeoperaattorilla merkitsevyystasolla 0.05. Erot olivat käytännössä olennaiset. Molempien operaattoreiden osalta tulos siis oli hypoteesin 4 mukainen.

Kun peruskyselyn tarkkuusarvoja verrattiin saman T5-tutkimusympäristön muiden kyselytyyppien tarkkuusarvoihin, olivat JA-operaattorin osalta erot verrattuna johdoskyselyyn ja yhdyssanakyselyyn tilastollisesti merkitseviä merkitsevyystasolla 0.05 ja ero yhdistelmäkyselyyn tilastollisesti merkitsevä

Taulukko 27. T1- ja T5-ympäristöissä kyselytyypeillä A, AB, AC ja ABC saatujen hakutulosten väliset erot johdososajoukossa (N = 8).

Ope-raattori	Kysely-tyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T5**	**T1**	**T5**	**T1**	**T5**
JA	A		12,3		51,9		67,3
	AB		27,1		74,5		56,5
	AC		18,3		58,2		54,9
	ABC	**41,3**	**41,5**	**82,7**	**88,8**	**46,9**	**47,8**
Virke	A		4,4		34,3		88,6
	AB		11,3		52,8		76,4
	AC		6,4		38,0		71,8
	ABC	**15,9**	**16,3**	**56,2**	**59,0**	**64,3**	**65,4**

A	Peruskysely
AB	Johdoskysely
AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat

merkitsevyystasolla 0.01. Virkeoperaattoria käytettäessä taas peruskyselyn tarkkuusarvo erosi sekä yhdyssana- että yhdistelmäkyselyjen tarkkuusarvoista merkitsevyystasolla 0.05. Kaikissa edellä mainituissa tapauksissa erot olivat Sparck Jonesin määrittelyn mukaan käytännössä olennaiset.

Johdososajoukossa siis peruskyselyn tulosjoukkojen saantiarvot jäivät huomattavasti alle muiden kyselytyyppien tulosjoukkojen saantiarvojen. Peruskyselyn tulosjoukkojen tarkkuusarvot kuitenkin olivat selvästi paremmat kuin muiden tulosjoukkojen tarkkuusarvot. Johdososajoukossa voidaan siis todeta, että mikäli haun tarkkuus on ehdottoman tärkeää, voidaan käyttää peruskyselyä - tosin sen kustannuksella, että saanti romahtaa.

Yhdyssanakysely

Yhdyssanakyselyn (AC) **saanti** oli JA-operaattorilla 25 ja virkeoperaattorilla 18 prosenttiyksikköä alempi kuin perinteisen yhdistelmäkyselyn (ABC). Molemmilla operaattoreilla saantiarvojen väliset erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.01 ja erot olivat käytännössä olennaiset (Sparck Jonesin mittari).

Verrattuna muihin T5-ympäristön kyselyihin yhdyssanakyselyn saanti oli vain peruskyselyn saantia parempi (merkitsevyystasolla 0.05, ero käytännössä huomattava). Muuten johdoskyselyn ja yhdistelmäkyselyn saantiarvot olivat yhdyssanakyselyn saantiarvoja paremmat - erot olivat tilastollisesti merkitsevät ja käytännössä olennaiset.

Yhdyssanakyselyn **tarkkuus** oli sekä JA- että virkeoperaattoria käytettäessä 8 prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn. Tarkkuusarvojen välinen ero oli JA-operaattorin tapauksessa tilastollisesti merkitsevä merkitsevyystasolla 0.01 ja käytännössä huomattava. Virkeoperaattoria käytettäessä tarkkuusarvojen välinen ero ei ollut tilastollisesti merkitsevä.

Kun yhdyssanakyselyn tarkkuusarvoja verrattiin T5-ympäristön johdososajoukon muiden kyselytyyppien tarkkuusarvojen kanssa, sen tarkkuusarvo oli JA-operaattoria käytettäessä alempi kuin peruskyselyn, mutta korkeampi kuin yhdistelmäkyselyn tarkkuusarvo - erot olivat tilastollisesti merkitseviä merkitsevyystasolla 0.05. Peruskyselyyn verrattaessa ero oli käytännössä olennainen, yhdistelmäkyselyyn verrattaessa käytännössä huomattava.

Eli johdososajoukossa peruskyselyn laajentaminen yhdyssanoilla paransi saantiarvoja jonkin verran, mutta samalla tarkkuus laski monta prosenttiyksikköä enemmän. Yhdyssanoihin laajentaminen ei siis ollut T5-ympäristön tässä osajoukossa käyttökelpoinen kyselyn laajentamiskeino.

Johdoskysely

T5-ympäristön johdoskyselyn (AB) suhteellinen **saanti** jäi JA-operaattorilla 8 ja virkeoperaattorilla 3 prosenttiyksikköä alemmaksi kuin perinteisen yhdistelmäkyselyn (ABC) saanti. Saantiarvojen välinen ero oli JA-operaattoria käytettäessä tilastollisesti merkitsevä merkitsevyystasolla 0.01. JA-operaattorilla ero oli myös käytännössä huomattava (Sparck Jonesin mittari). - Vastaavasti verrattaessa johdoskyselyä saman T5-ympäristön yhdistelmäkyselyn kanssa saantiarvojen välinen ero oli JA-operaattorilla tilastollisesti merkitsevä merkitsevyystasolla 0.01 ja käytännössä olennainen.

Virkeoperaattoria käytettäessä johdoskyselyn saantiarvo erosi sekä T1- että T5-ympäristön yhdistelmäkyselystä tilastollisesti merkitsevästi merkitsevyystasolla 0.05. Jälkimmäisessä tapauksessa eli T5-ympäristön sisäisessä

vertailussa ero oli käytännössä huomattava. Hypoteesi 5 ei siis saanut saannin osalta tukea.

Johdoskyselyn **tarkkuus** oli 10 (JA-operaattori) tai 12 (virkeoperaattori) prosenttiyksikköä korkeampi kuin perinteisen yhdistelmäkyselyn (ABC). Virkeoperaattorin tapauksessa tarkkuusarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.05 ja käytännössä olennainen (Sparck Jonesin mittari). - Sama tulos saatiin myös verrattaessa johdoskyselyä saman T5-ympäristön yhdistelmäkyselyyn.

Peruskyselyn laajentaminen johdoksilla siis auttoi johdososajoukossa nostamaan tulosjoukon saantiarvoja huomattavasti - ei kuitenkaan aivan perinteisen yhdistelmäkyselyn tasolle. Ero perinteiseen yhdistelmäkyselyyn oli silti pienempi kuin ero oman T5-ympäristön saantiarvoihin. Lisäksi johdosten lisääminen kasvatti saantia useita prosenttiyksikköjä enemmän kuin laski tarkkuutta.

Yhdistelmäkysely

Johdososajoukossa T5-ympäristön yhdistelmäkyselyjen (ABC) **tarkkuusarvo** oli kummankin operaattorin tapauksessa noin prosenttiyksikön verran korkeampi kuin T1-tutkimusympäristön yhdistelmäkyselyjen tarkkuusarvo, eli tutkimusympäristöjen välillä ei ollut sanottavaa eroa. Tämä tulos ei ole hypoteesin 7 mukainen.

Yhdistelmäkyselyjen **saantiarvoissa** erot olivat hiukan selvemmät: JA-operaattoria käytettäessä T5-ympäristön saantiarvot olivat 6 ja virkeoperaattoria käytettäessä 3 prosenttiyksikköä korkeammat kuin perinteisen yhdistelmäkyselyn. Saantiarvojen väliset erot eivät kuitenkaan olleet tilastollisesti merkitseviä. Tämäkään ei siis anna riittävästi tukea hypoteesille 7.

T5-ympäristössä johdososajoukon yhdistelmäkyselyllä **sekä** suhteellinen saanti **että** tarkkuus olivat paremmat kuin perinteisen yhdistelmäkyselyn. Tarkkuus oli parempi kuin hakijan katkaisemia hakusanoja käytettäessä, koska haut kohdistettiin täsmällisiin perusmuotoihin. Katkaistuja hakusanoja käytettäessä ei epärelevantteja dokumentteja pystytä välttämään yhtä hyvin. Tosin hyvä saanti edellyttää, että johdokset ja yhdyssanat muistetaan ottaa mukaan kyselyyn. Verrattuna perinteisen yhdistelmäkyselyyn erot eivät kuitenkaan olleet suuria.

9.5.3 Yhdyssanaosajoukko

Osittamattomien ja ositettujen hakusanojen väliset erot

Yhdyssanaosajoukossa oli 9 kyselyä, joissa yhdyssanat jaettiin osiinsa (taulukko 28; liite 14, luku 3). T5-ympäristössä yhdyssanojen osilla voitiin hakea niin yhdyssanojen alku-, keski- kuin loppuosina esiintyneitä sanoja.

Kaikkien kyselytyyppien, joissa yhdyssanoja ei jaettu osiinsa (siis T1/ABC, T5/A, T5/AB, T5/AC ja T5/ABC), **saantiarvot** erosivat tilastollisesti merkitsevästi merkitsevyystasolla 0.01 sekä T1- ja T5-ympäristöjen osien yhdistelmäkyselyn (ABCabc) että T5-ympäristön osien yhdyssanakyselyn (ACac) saantiarvoista. Tämä päti niin JA- kuin virkeoperaattoriakin käytettäessä.

Myös osien johdoskyselyn (ABab) saanti erosi tilastollisesti merkitsevästi osittamattomien kyselytyyppien saannista, mutta ero ei ollut yhtä selvä kuin edellisessä kappaleessa mainituilla kyselytyypeillä: JA-operaattoria käytettäessä saantiarvojen välinen ero oli merkitsevä merkitsevyystasolla 0.01, kun osien johdoskyselyn vertailukohteina olivat T5/A, T5/AB ja T5/AC. Edelleen erot olivat merkitseviä merkitsevyystasolla 0.05, kun sen vertailukohteina olivat yhdistelmäkyselyt (T1/ABC ja T5/ABC). Virkeoperaattoria käytettäessä taas osien johdoskyselyn saannin ja osittamattomien kyselytyyppien saannin väliset erot olivat yleensä merkitsevä merkitsevyystasolla 0.05 – poikkeuksena tästä peruskysely (A) ja yhdyssanakysely (AC), joiden tapauksessa tilastollinen merkitsevyys oli parempi eli merkitsevyystasolla 0.01.

Kun osittamattomia hakusanoja sisältävien kyselytyyppien saantiarvoja verrattiin keskenään, niiden välillä ei juuri ollut tilastollisesti merkitseviä eroja. Virkeoperaattoria käytettäessä näitä eroja ei ollut lainkaan. JA-operaattorin tapauksessa vertailuissa paljastui vain yksi osittamattomia yhdyssanoja sisältävien kyselytyyppien välinen ero: T5-ympäristön peruskyselyn (A) saanti erosi T1- ja T5-ympäristöjen yhdistelmäkyselyjen (ABC) saannista merkitsevästi merkitsevyystasolla 0.05. - Edellisen perusteella siis voidaan todeta, että yhdysanojen osittaminen kasvattaa tulosjoukon saantia huomattavasti.

Osittamattomia yhdyssanoja sisältävien kyselytyyppien **tarkkuusarvojenkaan** kesken ei juuri ollut tilastollisesti merkitseviä eroja. JA-operaattoria

Taulukko 28. T1- ja T5-ympäristöissä kaikilla kyselytyypeillä saatujen hakutulosten väliset erot yhdyssanaosajoukossa (N = 9).

Ope-raattori	Kysely-tyyppi	Tulosjoukon koko (keskim.)		Suhteellinen saanti %		Tarkkuus %	
		T1	**T5**	**T1**	**T5**	**T1**	**T5**
JA	A		6,1		48,8		86,1
	AB		6,7		51,2		84,7
	AC		6,8		53,4		80,1
	ABC	8,0	8,4	55,7	60,4	76,1	78,5
	Aa		11,9		67,4		65,7
	ABab		14,3		73,0		56,1
	ACac		20,7		84,4		51,1
	ABCabc	**25,8**	**25,0**	**88,1**	**98,4**	**42,1**	**46,6**
Virke	A		5,6		47,1		87,5
	AB		5,9		48,8		87,5
	AC		5,7		47,1		82,0
	ABC	6,1	6,9	48,8	52,9	80,0	81,1
	Aa		6,8		56,1		89,9
	ABab		7,2		58,5		90,2
	ACac		8,6		64,1		82,5
	ABCabc	**8,6**	**10,1**	**62,8**	**71,7**	**81,1**	**82,1**

A	Peruskysely
AB	Johdoskysely
AC	Yhdyssanakysely
ABC	Yhdistelmäkysely
Aa	Osien peruskysely
ABab	Osien johdoskysely
ACac	Osien yhdyssanakysely
ABCabc	Osien yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat

käytettäessä ainoastaan T5-ympäristön peruskyselyn tarkkuus oli perinteisen yhdistelmäkyselyn tarkkuutta parempi niin, että ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.05. Virkeoperaattoria käytettäessä tilastollisesti merkittäviä eroja ei löytynyt. Sen sijaan ositettuja hakusanoja sisältävien kyselytyyppien keskinäisessä vertailussa löytyi tilastollisesti merkitseviä eroja, joista tarkemmin seuraavissa alaluvuissa.

Saanniltaan parhaat kyselytyypit eli T1- ja T5-ympäristöjen osien yhdistelmäkyselyt (ABCabc) sekä osien yhdyssanakysely (T5/ACac) olivat tarkkuu-

deltaan huonoimmat (taulukko 28). Kun hakusanat oli kytketty toisiinsa JA-operaattorilla, näiden kyselytyyppien tarkkuusarvot erosivat toisten kyselytyyppien tarkkuusarvoista tavallisesti merkitsevyystasolla 0.01, parissa tapauksessa merkitsevyystasolla 0.05. Toisaalta virkeoperaattoria käytettäessä eri kyselytyyppien välillä ei ollut tilastollisesti merkitseviä eroja – kun yhdyssanat jaetaan osiinsa, kannattaa hakusanat siis kytkeä virkeoperaattorilla, koska silloin saanti paranee ilman että tarkkuus samassa määrin kärsii.

Osien peruskysely

Osien peruskyselyllä (Aa) saadun tulosjoukon **saanti** oli JA-operaattoria käytettäessä 21 ja virkeoperaattoria käytettäessä 7 prosenttiyksikköä alempi kuin perinteisen osien yhdistelmäkyselyn (ABCabc) tuottaman tulosjoukon saantiarvo. JA-operaattorin tapauksessa ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01 ja virkeoperaattoria käytettäessä merkitsevyystasolla 0.05. Käytännössä ero oli JA-operaattorin tapauksessa olennainen ja virkeoperaattorin tapauksessa huomattava (Sparck Jonesin mittari). Tulos tukee hypoteesia 4.

Toisaalta T5-ympäristön osien peruskyselyn (Aa) **tarkkuus** oli JA-operaattoria käytettäessä 24 ja virkeoperaattoria käytettäessä 9 prosenttiyksikköä korkeampi kuin perinteisen osien yhdistelmäkyselyn (ABCabc). Ero oli JA-operaattorilla merkitsevä merkitsevyystasolla 0.01, sen sijaan virkeoperaattorilla ero ei ollut tilastollisesti merkitsevä. Ero oli JA-operaattorilla käytännössä olennainen. JA -operaattorin osalta tulos on hypoteesin 4 mukainen, virkeoperaattorin osalta ei.

Kun osien peruskyselyä verrattiin saman T5-ympäristön muiden kyselytyyppien kanssa, saantiarvojen välillä todettiin tilastollisesti merkitsevä ero merkitsevyystasolla 0.01, kun peruskyselyn tulosjoukkoa verrattiin osien yhdyssanakyselyn (ACac) ja osien yhdistelmäkyselyn (ABCabc) tulosjoukkojen kanssa ja hakusanat oli kytketty JA-operaattorilla. Erot olivat käytännössä olennaiset. - Kun hakusanat oli kytketty virkeoperaattorilla, osien peruskyselyn ja osien yhdyssanakyselyn välinen ero oli merkitsevä merkitsevyystasolla 0.05 (käytännössä huomattava); osien peruskyselyn ja osien yhdistelmäkyselyn välinen ero oli merkitsevä merkitsevyystasolla 0.01 (käytännössä olennainen). Osien peruskyselyn ja osien johdoskyselyn välinen ero ei siis ollut kummallakaan operaattorilla tilastollisesti merkitsevä.

Tarkkuusarvoja verrattaessa T5-ympäristön osien peruskyselyn ja muiden ositettujen kyselytyyppien välille löytyi JA-operaattorilla seuraavat tilastollisesti merkitsevät erot: osien yhdistelmäkyselyn (ABCabc) kanssa merkitsevyystasolla 0.01, osien yhdyssanakyselyn (ACac) kanssa merkitsevyystasolla 0.05. Virkeoperaattoria käytettäessä ero eivät olleet tilastollisesti merkitseviä

Vaikka osien peruskyselyn tarkkuus siis olikin parempi kuin muilla kyselytyypeillä, sen saanti jäi selvästi huonommaksi kuin muiden yhdyssanojen osia sisältävien kyselytyyppien.

Osien johdoskysely

Kun osien peruskysely laajennettiin johdosperheellä osien johdoskyselyksi (ABab), tulosjoukkojen saanti kasvoi JA-operaattoria käytettäessä vajaat 6 prosenttiyksikköä ja virkeoperaattoria käytettäessä reilut 2 prosenttiyksikköä. Tällöin osien johdoskyselyn suhteellinen **saanti** oli JA-operaattoria käytettäessä 15 ja virkeoperaattoria käytettäessä 4 prosenttiyksikköä alempi kuin T1-ympäristön yhdistelmäkyselyn (ABCabc) tulosjoukon saantiarvo. JA-operaattorin tapauksessa osien johdoskyselyn ja osien yhdistelmäkyselyn saantiarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01 (ja käytännössä olennainen), virkeoperaattorin tapauksessa taas ero ei ollut tilastollisesti merkitsevä. Virkeoperaattoria käytettäessä tulos oli hypoteesin 5 mukainen, JA-operaattoria käytettäessä ei.

Osien johdoskyselyn tulosjoukon **tarkkuus** puolestaan oli JA-operaattorilla 14 ja virkeoperaattorilla 9 prosenttiyksikköä korkeampi kuin perinteisen osien yhdistelmäkyselyn tulosjoukon tarkkuus. JA-operaattorin tapauksessa tarkkuusarvojen välinen ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01, virkeoperaattorilla ero ei ollut tilastollisesti merkitsevä. Sparck Jonesin mittarin mukaan ero oli JA-operaattorin tapauksessa käytännössä olennainen. JA-operaattorin osalta tulos oli hypoteesin 5 mukainen, virkeoperaattorilla ei.

Kun osien johdoskyselyä verrattiin saman tutkimusympäristön muiden kyselytyyppien välillä, sen saantiarvot erosivat tilastollisesti merkitsevästi merkitsevyystasolla 0.01 sekä osien yhdyssanakyselyn että osien yhdistelmäkyselyn saantiarvoista, kun hakusanat oli yhdistetty JA-operaattorilla. Virkeoperaattoria käytettäessä osien johdoskyselyn saantiarvo erosi vain osien yhdistelmäkyselyn saantiarvosta, siitäkin merkitsevyystasolla 0.05.

Tarkkuusarvoja vertailtaessa todettiin, että osien johdoskyselyn tarkkuusarvo erosi tilastollisesti merkitsevästi vain osien yhdistelmäkyselyn tulosjoukkojen tarkkuusarvosta; siinäkin vain JA-operaattoria käytettäessä merkitsevyystasolla 0.05 (käytännössä huomattava) - virkeoperaattoria käytettäessä ei tilastollisesti merkitseviä eroja ollut.

Osien peruskyselyn laajentaminen johdosperheellä osien johdoskyselyksi ei siis tuota olennaista muutosta saanti- ja tarkkuusarvoihin, vaan muut laajennuskeinot ovat toimivampia.

Osien yhdyssanakysely

Osien yhdyssanakyselyn (ACac) suhteellinen **saanti** oli JA-operaattoria käytettäessä 4 prosenttiyksikköä alempi, mutta virkeoperaattoria käytettäessä reilun prosenttiyksikön verran *korkeampi* kuin perinteisen osien yhdistelmäkyselyn (ABCabc). Kummassakaan tapauksessa kyselytyyppien saantiarvojen välinen ero ei ollut tilastollisesti merkitsevä.

Osien yhdyssanakyselyn **tarkkuus** taas oli JA-operaattoria käytettäessä 9 prosenttiyksikköä ja virkeoperaattoria käytettäessä reilun prosenttiyksikön verran korkeampi kuin perinteisen osien yhdistelmäkyselyn. Tarkkuusarvojenkaan väliset erot eivät olleet tilastollisesti merkitseviä.

Pelkästään lisäämällä hakijan antamiin perusmuotoisiin hakusanoihin (eli ositettuihin yhdyssanoihin) yhdyssanojen eri osia symboloivat katkaisumerkit ja lisäämällä ne kyselyyn voidaan osien yhdyssanakyselyllä siis yltää lähes samaan saantiarvoon kuin perinteisellä tavalla haettaessa. Menetelmän etuna on, että se on helppo toteuttaa automaattisesti.

Osien yhdistelmäkysely

T5-ympäristön osien yhdistelmäkyselyn (ABCabc) **tarkkuus** oli JA-operaattoria käytettäessä vajaat 5 ja virkeoperaattoria käytettäessä yhden prosenttiyksikön verran korkeampi kuin perinteisen osien yhdistelmäkyselyn tarkkuus. Tulos ei ole hypoteesin 7 mukainen, koska sen mukaan tarkkuuden piti olla keskimäärin huonompi kuin käytettäessä hakijan katkaisemia hakusanoja; nyt ero ei ollut tilastollisesti merkitsevä. Tosin tulos siinä mielessä on mielenkiintoinen, että sen perusteella keski- ja loppuosien lisääminen ei siis huononnakaan hakutuloksen tarkkuutta.

T5-ympäristö oli myös **saantiarvoiltaan** T1-ympäristöä **parempi**: JA-operaattoria käytettäessä sen osien yhdistelmäkyselyn saantiarvot olivat 10 ja virkeoperaattoria käytettäessä 9 prosenttiyksikköä korkeammat kuin vastaavan kyselytyypin saantiarvot T1-ympäristössä. Edellämainitut saanti- ja tarkkuusarvojen väliset erot eivät kuitenkaan olleet tilastollisesti merkitseviä, joten hypoteesi 7 ei tässäkään saanut riittävästi tukea.

Yhteenveto

Monissa tapauksissa yhdyssanojen keski- ja loppuosien lisäämisellä ei ollut hakutulosten kannalta merkitystä, koska hakusana ei esiintynyt kuin yhdyssanan alussa. Joissain tapauksissa, kuten etsittäessä artikkeleita Wärtsilän tilintarkastuksesta (hakupyyntö 20), dokumenttien määrä kuitenkin kasvoi huomattavasti verrattuna kyselyyn, joka tehdään taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla. T5-ympäristössä tulosjoukkoihin saatiin muun muassa sellaisia relevantteja dokumentteja, joissa ei ollut esiintynyt yhtään tilintarkastus-alkuista sananmuotoa, mutta ne tulivat tulosjoukkoon siksi, että niissä esiintyi sana valtiontilintarkastaja.

T5-ympäristön tulosten perusteella voidaan todeta, että yhdyssanojen keski- ja loppuosilla hakeminen (mikä ei muissa tutkimusympäristöissä käytännössä ollut mahdollista), nostaa tulosjoukon saantiarvoa ilman, että tarkkuusarvo vastaavassa määrin alenisi. Käytännössä tarkkuus pysyi vähintään samalla tasolla kuin perinteisellä tavalla haettaessa. Jos siis hakujen saanti halutaan mahdollisimman hyväksi, tulisi myös yhdyssanojen keski- ja loppuosat tallentaa hakemistoon.

Miksi sitten yhdyssanojen osien lisääminen kyselyyn ei erityisesti huononna tarkkuutta? Yksi selitys on se, että jos kyselyssä on useampi kuin yksi hakusana, tällaiset rajaavat käsitteet auttavat karsimaan väärät osumat pois (erityisesti silloin, kun hakusanat ovat kytketty toisiinsa suhteellisen tiiviillä eli JA-operaattoria tiukemmalla operaattorilla). Vaikka hakusana yksinään tuottaisikin paljon osumia, kyselyn muut hakusanat rajoittavat haun alaa tehokkaasti (Lancaster 1986, s. 69). Näinhän käy silloin, kun yhdyssana jaetaan: siitä syntyy vähintään kaksi eri hakusanaa (eli rajaavaa ilmausta), jotka kytketään toisiinsa rajaavalla operaattorilla.

Jos T5-ympäristössä halutaan saada hakutulos, jonka tarkkuusarvo on mahdollisimman korkea, tämä käy päinsä käyttämällä kyselyssä vain hakusanan perusmuotoja. Tällaisen peruskyselyn saanti on kuitenkin huono verrattuna

perinteiseen yhdistelmäkyselyyn. Sitä paitsi saanti laskee useampia prosenttiyksikköjä enemmän kuin tarkkuus paranee.

Kun hakija haluaa hyvän saannin ja lisää kyselyihin hakusanojen rinnalle yhdyssanojen osat (osien yhdyssanakysely), saanti on jokseenkin yhtä hyvä kuin perinteisen osien yhdistelmäkyselyn ja tarkkuus tätä korkeampi. Eli jo tällä tavalla ylletään samaan tulokseen kuin perinteisellä osien yhdistelmäkyselyllä. Sitä parempi saanti saadaan vielä, kun kyselyyn lisätään myös johdosperhe (yhdyssanojen osat ja näiden johdosperhe yhdyssanan osina). T5-ympäristön (osien) yhdistelmäkyselyllä sekä saanti että tarkkuus ovat korkeammat kuin vastaavalla perinteisellä kyselytyypillä.

9.6 Eri tutkimusympäristöjen vertailu keskenään

Muut tutkimusympäristöt voidaan karkeasti suhteuttaa T1-tutkimusympäristöön seuraavasti: T2-, T3- ja T4-ympäristöissä pyrittiin saamaan perinteistä hakutapaa parempi **tarkkuus** samalla, kun saanti olisi suunnilleen sen kanssa samalla tasolla. Erityisesti T3-ympäristössä pyrittiin saamaan hyvä tarkkuus seulomalla väärät osumat pois. T5-ympäristössä pyrittiin erityisesti saamaan parempi **saanti** kuin T1-ympäristössä, koska siinä oli mahdollista laajentaa kyselyä myös yhdyssanan keski- ja loppuosilla. Samalla toki pyrittiin siihen, ettei tarkkuus olennaisesti heikkenisi.

Eri tutkimusympäristöjä verrattiin toisiinsa niissä kaikissa olleiden kyselytyyppien eli yhdistelmäkyselyn (ABC) ja toisaalta osien yhdistelmäkyselyn (ABCabc) perusteella. Näiden laajimpien kyselytyyppien tuottamien tulosjoukkojen koot olivat eri tutkimusympäristöissä hyvin lähellä toisiaan; tästä poikkeuksena oli vain T3-ympäristö eli seulonta, jossa tulosjoukkojen koko oli selvästi muita pienempi. Vertailuissa käytettiin Friedmanin merkitsevyystestiä (liite 14, luku 4).

9.6.1. Perusjoukko

Perusjoukkoon sisältyi yhdistelmäkyselyjen vertailussa 25 kyselyä (taulukko 29). Vertailuista jätettiin pois kysely 28, jonka tarkkuusarvoa ei T3-ympäristössä voitu laskea (ks. tarkemmin luku 9.3).

Taulukko 29. Eri tutkimusympäristöjen yhdistelmäkyselyn vertailu perusjoukossa (N = 25).

S/T	Ope-raattori	Kysely-tyyppi	**T1**	**T2**	**T3**	**T4**	**T5**
Suht. saanti %	JA	ABC	73,8	73,6	70,0	73,8	**75,6**
	Virke	ABC	61,8	61,6	59,1	61,9	**62,8**
Tarkkuus %	JA	ABC	68,4	69,1	69,8	**70,7**	**70,7**
	Virke	ABC	77,0	77,0	**78,4**	77,8	77,7

ABC	Yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat
T3	Taivutusmuotohakemisto, seulotut hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat

Perusjoukossa T3-ympäristön yhdistelmäkyselyn **saanti** jäi huonommaksi kuin muiden tutkimusympäristöjen saantiarvot. Ero muiden tutkimusympäristöjen saman kyselytyypin välillä oli tilastollisesti merkitsevä, JA-operaattoria käytettäessä merkitsevyystasolla 0.025 ja virkeoperaattoria käytettäessä merkitsevyystasolla 0.01. Muita tilastollisesti merkitseviä eroja eri tutkimusympäristöjen välillä ei ollut.

Eri tutkimusympäristöjen tulosjoukkojen **tarkkuusarvojen** välillä ei ollut tilastollisesti merkitseviä eroja.

9.6.2. Johdososajoukko

Myös johdososajoukossa (taulukko 30) T3-ympäristön yhdistelmäkyselyn tulosjoukon **saanti** jäi alemmaksi kuin muilla vertailluilla tulosjoukoilla. JA-operaattoria käytettäessä ero ei varsinaisesti ollut tilastollisesti merkitsevä (vasta merkitsevyystasolla 0.1). Virkeoperaattorilla ero oli selvempi: se oli tilastollisesti merkitsevä merkitsevyystasolla 0.025 (ja käytännössä huomattava). Muita tilastollisesti merkitseviä eroja eri tutkimusympäristöjen saantiarvojen välillä ei ollut.

Yhdistelmäkyselyjen **tarkkuusarvojen** väliset erot eivät olleet johdososajoukossa tilastollisesti merkitseviä.

Taulukko 30. Eri tutkimusympäristöjen yhdistelmäkyselyn vertailu johdososajoukossa (N = 8).

S/T	Operaattori	Kysely-tyyppi	**T1**	**T2**	**T3**	**T4**	**T5**
Suht. saanti %	JA	ABC	82,7	82,7	73,7	83,4	**88,8**
	Virke	ABC	56,2	56,2	49,9	56,9	**59,0**
Tarkkuus %	JA	ABC	46,9	46,9	**48,3**	48,2	47,8
	Virke	ABC	64,3	64,3	**67,9**	66,8	65,4

ABC	Yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat
T3	Taivutusmuotohakemisto, seulotut hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat

9.6.3. Yhdyssanaosajoukko

Yhdyssanaosajoukossa (taulukko 31) oli T3-ympäristön osien yhdistelmäkyselyn tulosjoukon **saantiarvo** muiden tutkimusympäristöjen saantiarvoja huonompi. JA-operaattoria käytettäessä T3-ympäristön saanti erosi kaikkien muiden tutkimusympäristöjen vastaavista arvoista tilastollisesti merkitsevästi merkitsevyystasolla 0.01.

T5-ympäristön osien yhdistelmäkyselyn saantiarvot olivat korkeammat kuin toisissa tutkimusympäristöissä. Ero ei kuitenkaan ollut niin suuri, että se olisi ollut selvästi tilastollisesti merkitsevä. Verrattuna T1- ja T2-ympäristöjen saantiarvoihin erot tosin olivat tilastollisesti merkitsevät merkitsevyystasolla 0.1, mutta sitä ei voi pitää riittävänä.

Virkeoperaattorin tapauksessa löytyi toisenlainen ero: T5-ympäristön tulosjoukon saantiarvo oli selvästi parempi kuin toisissa tutkimusympäristöissä. Se erosi T1-, T2- ja T3-ympäristöjen tulosjoukkojen saantiarvosta tilastollisesti merkitsevästi merkitsevyystasolla 0.025. Lisäksi T5-ympäristön saantiarvot olivat T4-ympäristön saantiarvoja paremmat merkitsevyystasolla 0.05. Kaikissa edellämainituissa tapauksissa erot olivat Sparck Jonesin määrittelyjen mukaan käytännössä huomattavat.

Yhdyssanaosajoukon **tarkkuusarvoja** vertailtaessa todettiin, että JA-operaattoria käytettäessä T3-ympäristön osien yhdistelmäkyselyn tarkkuus oli

Taulukko 31. Eri tutkimusympäristöjen (osien) yhdistelmäkyselyn vertailu yhdysanaosajoukossa (N = 9).

S/T	Ope-raattori	Kysely-tyyppi	**T1**	**T2**	**T3**	**T4**	**T5**
Suhteel-linen saanti %	JA	ABC	55,7	55,7	54,5	56,3	60,4
		ABCabc	88,1	88,1	83,5	88,6	**98,4**
	Virke	ABC	48,8	48,8	48,3	49,4	52,9
		ABCabc	62,8	62,8	62,2	63,4	**71,7**
Tarkkuus %	JA	ABC	76,1	78,2	77,7	79,4	78,5
		ABCabc	42,1	43,0	**50,4**	44,6	46,6
	Virke	ABC	80,0	80,0	79,8	82,3	81,1
		ABCabc	81,1	81,4	81,2	**83,7**	82,1

ABC	Yhdistelmäkysely
ABCabc	Osien yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat
T3	Taivutusmuotohakemisto, seulotut hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat

selvästi T1- ja T2-ympäristöjen vastaavan kyselyn tarkkuutta parempi. Tämä ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.01. Lisäksi T3-ympäristön osien yhdistelmäkyselyn tarkkuusarvo oli vastaavaa T4-ympäristön kyselyn tarkkuusarvoa parempi niin, että ero oli tilastollisesti merkitsevä merkitsevyystasolla 0.05. Kaikissa edellämainituissa tapauksissa erot olivat myös käytännössä huomattavat. Sen sijaan verrattaessa T3-ympäristön osien yhdistelmäkyselyä T5-ympäristön osien yhdistelmäkyselyyn ero oli merkitsevä vasta merkitsevyystasolla 0.1; tuota merkitsevyystasoa ei voi pitää riittävänä.

Myös T4- ja T5-ympäristöjen tarkkuusarvot olivat T1-ympäristön tarkkuusarvoja korkeammat niin, että erot olivat tilastollisesti merkitsevät merkitsevyystasolla 0.05. Tässä tapauksessa erot eivät kuitenkaan Sparck Jonesin määrittelyn mukaan olleet käytännössä merkitsevät.

Virkeoperaattoria käytettäessä ei eri ympäristöjen tarkkuusarvojen välillä ollut tilastollisesti merkitseviä eroja.

9.6.4. Yhteenveto

Edellä esiteltyjen merkitsevyystestien selkein tulos oli, että T3-ympäristö eli seulonta osoittautui saannin suhteen huonoimmaksi. Se ei tarkkuudeltaankaan ollut ylivoimainen muihin verrattuna; vain yhdyssanaosajoukossa, JA-operaattoria käytettäessä, sen tarkkuusarvot olivat korkeammat kuin T1-, T2- ja T4-ympäristöissä. Samalla T3-ympäristön tarkkuusarvojen ero oli T5-ympäristöön verrattuna pienempi kuin muihin tutkimusympäristöihin verrattuna.

Jos eri tutkimusympäristöistä saatavien hakutulosten välillä ei ole suurta eroa, parhaana vaihtoehdoista voidaan pitää sitä, jossa kyselyt on helpointa toteuttaa. FULLTEXT-tutkimuksessa tutkimusympäristöjä ei testattu todellisten tiedontarvitsijoiden avulla, mutta on todennäköistä, että perusmuotoisten hakusanojen käyttö on ainakin satunnaiselle käyttäjälle helpompaa kuin hakusanojen katkaiseminen. Tosin satunnaista (ja miksei ammattilaistakin) hakijaa on opastettava lisäämään kyselyyn hakusanan rinnalle myös johdosperhe ja yhdyssanat, jotta saanti olisi hyvä, tai tällaista laajentamista varten tulee kehittää automaattisia välineitä.

Toisaalta on huomattava, että eri tutkimusympäristöt tarjoavat erilaiset mahdollisuudet haun variointiin. Esimerkiksi T4- ja T5-ympäristöissä suppean (osien) peruskyselyn ja laajimman (osien) yhdistelmäkyselyn tulosjoukkojen saanti- ja tarkkuusarvojen väliset erot olivat suuret; siten hakija voi niissä valita painottaako saantia vai tarkkuutta - T1-ympäristössä yhtä joustava vaihtelu ei ole mahdollinen.

Vaikka eri tutkimusympäristöjen välille merkitsevyystestien avulla löytyikin tiettyjä eroja, tilastollisesti merkitsevien erojen lisäksi voidaan tarkastella eri tutkimusympäristöjen ja kyselytyyppien välisten erojen johdonmukaisuutta. Keen (1992) totesi, että oma merkityksensä on sillä, jos tuloksissa näkyy selvä tendenssi niin, että toinen (tai jokin) vertailtavista ilmiöistä saa toista (tai muita) parempia arvoja, olkoonkin että erot olisivat pieniä. Kun eri tutkimusympäristöjen yhdistelmäkyselyjen ja osien yhdistelmäkyselyjen saanti- ja tarkkuusarvoja verrataan toisiinsa, eri tutkimusympäristöt voidaan asettaa karkeaan paremmuusjärjestykseen.

Yksi keino mitata tätä asiaa on hyödyntää Friedmanin merkitsevyystesteissä käytettyä asetelmaa, jossa kukin saanti- tai tarkkuusarvo korvattiin rangillaan eli järjestysluvullaan. Tällä tavalla voidaan tutkia, mikä vertailtavista

Taulukko 32. Eri tutkimusympäristöjen (osien) yhdistelmäkyselyn saantiarvojen vertailu Friedmanin testin mukaisen rankkeerauksen mukaan ja keskiarvojen perusteella. Paras saantiarvo on vasemmalla, huonoin oikealla.

Joukko	Operaattori	Friedman (ranki)	Keskiarvo
Perus (ABC)	JA	T1,T4,T5 > T2 > T3	T5 > T4,T1 > T2 > T3
	Virke	T1, T4 > T2 > T5 > T3	T5 > T4 > T1 > T2 > T3
Johdos (ABC)	JA	T4,T5 > T1,T2 > T3	T5 > T4 > T1,T2 > T3
	Virke	T4 > T1,T2 > T5 > T3	T5 > T4 > T1,T2 > T3
Yhdyss. (ABCabc)	JA	T5 > T4 > T1,T2 > T3	T5 > T4 > T1,T2 > T3
	Virke	T5 > T4 > T1,T2 > T3	T5 > T4 > T1,T2 > T3
Yleisarvio:		T4 > T5 > T1 > T2 > T3	T5 > T4 > T1 > T2 > T3

ABC	Yhdistelmäkysely
ABCabc	Osien yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat
T3	Taivutusmuotohakemisto, seulotut hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat

tutkimusympäristöistä on saanut korkeimman arvon, mikä toiseksi korkeimman ja niin edelleen, ja siten nähdä tendenssit. Taulukkojen 32 ja 33 Friedman-sarakkeessa esitetään eri tutkimusympäristöjen (osien) yhdistelmäkyselyjen saanti- ja tarkkuusarvojen keskinäinen järjestys. Alimpana taulukossa on vielä yhteenvetorivi, jossa eri tutkimusjoukkojen välisestä järjestyksestä on vielä tehty yhteenvetoanalyysi (synteesi).

Friedmanin testin rangeilla tutkitaan vain vertailtavien kohteiden järjestystä. Siinä ei siis ota kantaa erojen suuruuteen, vaan systemaattisuuteen: mikä vertailluista vaihtoehdoista on useimmin ollut paras. Muuten vaihtoehtojen väliset erot voivat olla hyvinkin pieniä. Erojen suuruutta taas voidaan mitata keskiarvolukujen avulla. Systemaattisuuden suhteen keskiarvo ei ole paras mahdollinen mittari, koska sitä voivat yksittäiset luvut muuttaa paljonkin. Näiden kahden eri mittarin tuloksia onkin mielenkiintoista verrata toisiinsa.

Saantiarvojen osalta parhaat tulokset saadaan T5- ja T4-ympäristöissä. Keskiarvojen perusteella T5-ympäristön joissain kyselyissä saatiin selvästi korkeammat saantiarvot kuin muissa tutkimusympäristöissä. Toisaalta taas T4-

Taulukko 33. Eri tutkimusympäristöjen (osien) yhdistelmäkyselyn tarkkuusarvojen vertailu Friedmanin testin rankkeerauksen mukaan ja keskiarvojen perusteella. Paras tarkkuusarvo on vasemmalla, huonoin oikealla.

Joukko	Operaattori	Friedman (ranki)	Keskiarvo
Perus (ABC)	JA	T4 > T5 > T3 > T2 > T1	T4, T5 > T3 > T2 > T1
	Virke	T4 > T3 > T2,T1 > T5	T3 > T4 > T5 > T1,T2
Johdos (ABC)	JA	T3,T4 > T5 > T1,T2	T3 > T4 > T5 > T1,T2
	Virke	T3 > T4 > T1,T2 > T5	T3 > T4 > T5 > T1,T2
Yhdyss. (ABCabc)	JA	T3 > T5 > T4 > T2 > T1	T3 > T5 > T4 > T2 > T1
	Virke	T4 > T2 > T5 > T1 > T3	T4 > T5 > T2 > T3 > T1
Yleisarvio:		T4 > T3 > T5 > T2 > T1	T3 > T4 > T5 > T2 > T1

ABC	Yhdistelmäkysely
ABCabc	Osien yhdistelmäkysely
T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat
T3	Taivutusmuotohakemisto, seulotut hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat

ympäristö oli systemaattisemmin paras eli siinä saatiin useimmiten paras saanti, joskaan ero muihin tutkimusympäristöihin ei välttämättä ollut suuri.

Muiden tutkimusympäristöjen keskinäinen järjestys on selvä, molemmat vertailutavat asettivat vaihtoehdot samaan järjestykseen. Tosin T1- ja T2-ympäristöjen välinen ero oli hiuksenhieno - monissa tapauksissa niiden saantiarvot olivat täsmälleen samat. Selvästi huonoimmat saantiarvot saatiin tutkimusympäristössä T3.

Vastaavantyyppinen tulos saatiin myös tarkkuusarvojen osalta: T4-ympäristö oli systemaattisemmin paras eli siinä saatiin useimmin parhaat tarkkuusarvot - joskin erot muiden ympäristöjen tarkkuusarvoihin saattoivat määrällisesti mitaten olla pienet. Toisaalta taas T3-ympäristössä saatiin yksittäisissä kyselyissä selvästi suuremmat tarkkuusarvot kuin muissa tutkimusympäristöissä, mikä siten nosti seulonnan keskiarvolla mitaten parhaaksi vaihtoehdoksi.

T5-ympäristö jäi tarkkuusvertailuissa kolmanneksi; tosin sen tarkkuusarvot olivat usein varsin lähellä T4- tai T3-ympäristön tarkkuusarvoja. Myös T1-

ja T2-ympäristöjen tarkkuusarvot olivat hyvin lähellä toisiaan, joskin näistä kahdesta T1-ympäristö jäi lopulta niukasti huonommaksi.

Mistä edellämainitut erot sitten johtuvat? Mitkä dokumentit esiintyvät jonkin tietyn tutkimusympäristön tulosjoukoissa ja mitkä niistä jäävät pois, kun eri tutkimusympäristöjen tulosjoukkoja verrataan keskenään?

T1- ja T2-ympäristöjen yhdistelmäkyselyjen ja osien yhdistelmäkyselyjen tulosjoukot poikkesivat toisistaan muutamissa tapauksissa. Erot johtuivat siitä, että hakijan katkaisema hakusana oli niin lyhyt, että se palautti hakemistosta myös muiden sanojen esiintymiä, kun taas vartalo-ohjelmien pidemmät vartalot olivat karsineet tällaiset pois. Esimerkiksi hakupyynnössä 19 ollut vientituki-sana katkaistiin T1-ympäristössä muotoon *vientitu**, joka palautti hakemistosta dokumentin, jossa esiintyi sana vientituotteita. Sen sijaan automaattisesti katkaistut vartalot olivat pidempiä ja eivätkä täsmänneet tämän sanan kanssa.

T3-ympäristössä hakemistosanat seulottiin. Seulonnan jälkeen haun tuloksena saatiin pienempi tulosjoukko, koska seulontamenetelmä ei hyväksynyt dokumentteja, joissa hakusana ei ollut täsmälleen sama kuin alkuperäinen hakusana. Tämä tarkoitti sitä, että myös hakusanalla alkavat yhdyssanat karsittiin pois. Rajauksesta seurasi, että myös relevantteja dokumentteja karsiutui pois ja saanti siten aleni perinteiseen yhdistelmäkyselyyn verrattuna. Seulonnan kautta muodostetun kyselyn tarkkuus nousi selvästi vastaavan perinteisen kyselyn tarkkuutta paremmaksi, mutta muiden ympäristöjen ja T3-ympäristön tarkkuusarvojen välinen ero ei ollut niin suuri.

T1- ja T4-ympäristöjen välillä erot johtuivat esimerkiksi siitä, että perusmuotohakemistossa käytettiin katkaisematonta hakusanaa *neste*, kun haettiin Neste-yhtiötä (hakupyyntö 24). Katkaistu hakusana *neste** taas palautti hakemistosta Neste-yhtiön taivutusmuotojen lisäksi myös yhdyssanan nestekaasu, joka tässä tapauksessa ei ollut relevantti. Toisaalta samasta syystä hakupyynnön 26 kyselyistä jäivät T4-ympäristön tulosjoukosta pois dokumentit, joissa esiintyi jokin japanilainen-adjektiivin muodoista. Nämä muodot saatiin taivutusmuotohakemistosta katkaistulla hakusanalla *japani**, kun taas perusmuotohakemistosta haettiin vain pelkällä perusmuotoisella hakusanalla *japani*.

T5-ympäristön erot muihin ympäristöihin nähden johtuivat siitä, että siinä tulosjoukkoihin saatiin dokumentit, joissa hakusana oli esiintynyt yhdys-

sanan keski- tai loppuosana. Esimerkiksi kyselyssä 6 löydettiin dokumentti (tosin epärelevantti sellainen), jossa hakusana tutkimus oli esiintynyt sanan-muodossa harjuntutkimusyksiköltä.

Vaikka tutkimustulosten analyysissä onkin paljon käsitelty johdoksia ja niiden mukaantulon ja poisjäännin vaikutuksia, on huomattava, että johdokset ovat vain yksi osa kyselyn laajentamista hakusanan rinnakkaisilmauksilla. Todellisessa hakutilanteessa on otettava huomioon myös synonyymit ja muut vastaavat ilmaukset. Tässä on käsitelty johdoksia, koska ne ovat morfologisen tason ilmiöitä ja siten kuuluvat tämän tutkimuksen alaan.

Varsinainen johdososajoukko oli melko pieni, vain kahdeksan kyselyä. Siten erot voisivat johtua pelkästään sattumasta eli johdoskyselyissa käytetyn joukon pienuudesta. Eroille voi kuitenkin esittää myös kielitieteellisen selityksen: jos sanalla on suuri johdosperhe (esimerkiksi tutkia, tutkiminen, tutkimus, tutkinta), sen taustalla olevan käsitteen merkityskenttä hajoaa usean sanan kesken. Johdosperheen yksittäinen jäsen siis kattaa vain osan käsitteen koko merkityskentästä. Tällöin tekstien kirjoittajilla on paljon vaihtoehtoja, miten ilmaista haluamansa asia (tutkia pohjavesiä, pohjavesien tutkiminen jne.). Mikäli hakusanana käytetään vain johdosperheen yhtä jäsentä, muut vaihtoehdot jäävät löytymättä ja saanti kärsii. Mikäli taas sanalla on vain vähän tai ei lainkaan johdoksia (esimerkiksi lomaosake), kirjoittajilla ei juuri ole valinnanvaraa, vaan heidän on käytettävä tätä ilmausta hakusanana. Vastaavasti hakutilanteessa saadaan yhdellä hakusanalla helposti katettua käsitteen koko merkityskenttä ja haun saanti on helppo saada korkeaksi.

Edellä kuvattu periaate pätee myös synonyymeillä, yhdyssanoilla ja muilla vastaavilla ilmiöillä, joissa käsitteelle ei ole vain yhtä mahdollista ilmausta: Jos sanalla on synonyymejä, merkityskenttä hajoaa usean vaihtoehdon kesken. Jos hakija käyttää vain yhtä synonyymiä, saanti kärsii. Jos taas hakusana on yhdyssana, joka dokumenteissa on esiintynyt myös sanaliittona (autoverotus - autojen verotus), saanti kärsii, jos kaikkia vaihtoehtoja ei oteta kyselyyn mukaan.

10 ONGELMAKYSELYT

Ongelmakyselyjen tarkoituksena oli selvittää, voidaanko hakemistosta tulkintaohjelmien avulla - tai niistä huolimatta - löytää juuri halutut sanat eli pystytäänkö kysely kohdistamaan täsmälleen tiettyihin sanoihin tai sananmuotoihin. Ongelmakyselyjä oli viisitoista: kyselyt 31 - 45. Näistä viidestätoista kahdeksan oli laadittu aidoista sanomalehtiarkiston hakupyynnöistä ja loput olivat tutkijan itse keksimiä.

Ongelmakyselyjen suhteellista saantia ja tarkkuutta ei määritelty aihe- tai hyötyrelevanssin perusteella, koska tämä ei olisi antanut riittävän täsmällistä tietoa. Tiedontarvitsijan kannalta on normaalisti samantekevää, esiintyykö relevantissa dokumentissa autoverotus vai autovero, mutta morfologisen tulkintaohjelman on toimittava täsmällisesti ja pystyttävä erottamaan nämä sanat toisistaan. Niinpä ongelmakyselyissä tulosten oikeellisuus tarkoitti morfologista hyväksyttävyyttä. Tämä määriteltiin seuraavasti: jos haluttu hakusana tai sanaliitto esiintyi dokumentissa, tämä dokumentti oli hyväksyttävä, muussa tapauksessa ei. Oikeellisuusanalyysin teki tutkija itse. (Analyysin tulokset kysymyksittäin liitteessä 15.)

Eri tutkimusympäristöissä saatuja tulosjoukkoja verrattiin toisiinsa ja niistä tutkittiin muun muassa, saatiinko perusmuotoisilla hakusanoilla samat dokumentit kuin hakijan katkaisemilla hakusanoilla. Lisäksi tutkittiin, sisälsivätkö tulosdokumentit todella hakusanan jonkin esiintymän vai joutuivatko ne tulosjoukkoon liian lyhyeksi katkaistun hakusanan, sananmuotohomografin väärän tulkinnan tms. vuoksi.

Ongelmakyselyt suoritettiin eri tutkimusympäristöissä eri tavoin hakemiston ominaisuuksien mukaan:

1. Hakijan katkaisemilla hakusanoilla taivutusmuotohakemistosta (T1)
2. Automaattisesti katkaistuilla (vartalo-ohjelman katkaisemilla) hakusanoilla taivutusmuotohakemistosta (T2)
3. Perusmuotoisilla hakusanoilla perusmuotohakemistosta sekä automaattisesti katkaistuilla hakusanoilla perusmuotohakemistosta ja tunnistamattomien sananmuotojen hakemistosta (T4)
4. Perusmuotoisilla hakusanoilla ositetusta perusmuotohakemistosta sekä automaattisesti katkaistuilla hakusanoilla tunnistamattomien sananmuotojen hakemistosta (T5)

5. Katkaistuilla hakusanoilla taivutusmuotohakemistosta, mikäli kysely perusmuotoisilla hakusanoilla ei tuottanut tulosta (T6).

Sen sijaan tutkimusympäristöstä T3 (seulonta) ongelmakyselyjä ei suoritettu. Tähän ympäristöön olisi tarvittu varsin monivaiheinen virheenkorjausprosessi, jonka ohjelmointi olisi vaatinut liikaa resursseja suhteessa projektin tavoitteisiin ja resursseihin.

Kun morfologisia tulkintaohjelmia sovelletaan hakujärjestelmissä, ongelmakyselyissä voi käydä niin, että tulosjoukon saanti tai tarkkuus jää morfologisten ohjelmien tulkintavirheiden takia huonommaksi kuin muuten vastaavan, mutta hakijan katkaisemilla hakusanoilla taivutusmuotohakemistosta saadun tulosjoukon. Tämä on otettava hakujärjestelmän toteutuksessa huomioon ja kehitettävä hyvän saannin ja tarkkuuden varmistamiseksi (mahdollisimman) automaattiset virheenkorjausmenetelmät.

Mikäli hakusanaa ei syötetä perusmuodossa tai sen sanaluokka ilmoitetaan taivutusvartalo-ohjelmalle väärin, vartalo-ohjelma ei tuota kaikkia tarvittavia vartaloita tai vartalot ovat virheellisiä. Perusmuoto-ohjelma puolestaan voi tulkita homografisen sananmuodon jonkin toisen, väärän sanan esiintymäksi, jos oikea perusmuoto sattuu puuttumaan perusmuoto-ohjelman sanakirjasta. Edelläkuvattujen virhetyyppien vuoksi osa hakusanan taivutusmuodoista voi jäädä löytymättä, minkä seurauksena **saanti** huononee.

Toisaalta taivutusvartalo-ohjelmat tuottavat tiettyjen taivutusluokkien sanoille varmuuden vuoksi ylimääräisiä vartaloita (vrt. rakkaus : rakkaude*, mutta toisaalta myös pakkaus - pakkaude*). Perusmuotoistettaessa taas perusmuoto-ohjelma voi löytää homografiselle ilmaukselle monta perusmuototulkintaa, jotka kaikki on tallennettava hakemistoon, vaikka vain yksi tulkinnoista on oikea. Näissä tapauksissa tulosjoukkoon voi joutua ylimääräisiä dokumentteja, mikä laskee tuloksen **tarkkuutta**.

Perusmuotohakemistosta ei voi suoraan etsiä hakusanan tiettyä taivutusmuotoa. Taivutusmuotohakemistossahan kysely voidaan kohdistaa juuri haluttuun taivutusmuotoon - esimerkiksi silloin, kun etsitään tietynnimistä kaunokirjallista teosta (Hänen olivat linnut, Täällä Pohjantähden alla). Mikäli perusmuotohakemiston avulla halutaan löytää tietty sananmuoto, tarvitaan välivaihe, jossa kysely ensin toteutetaan hakusanojen perusmuodoilla; tämän jälkeen tulosjoukosta karsitaan pois muut kuin ne dokumentit, joissa esiintyy haluttu taivutusmuoto. Kaikki muut taivutusmuodothan ovat

hakijan kannalta ylimääräisiä ja vain tuottavat tulosjoukkoon epärelevantteja dokumentteja, eli alentavat haun tarkkuutta. Tällaisen poiminnan tai karsinnan tulisi tapahtua automaattisesti niin, että hakija pystyisi helposti määrittelemään haluamansa taivutusmuodon tai -muodot, minkä jälkeen hakujärjestelmä huolehtisi oikeiden sananmuotojen löytämisestä.

10.1 Ongelmakyselyjen tyypit

Perusmuotoistamiselle ongelmia tuottavat hakusanat voidaan ryhmitellä seuraavasti:

1) Hakusana ei löydy perusmuotoon palauttavan ohjelman sanakirjasta eikä sitä (perusmuodossaan) voida tulkita minkään tunnetun sanan taivutusmuodoksi.
2) Hakusana sellaisenaan ei löydy sanakirjasta, mutta se voidaan tulkita sanakirjassa olevan sanan taivutusmuodoksi. Joko
    2a) kyseessä todella on tuntematon perusmuoto, jonka perusmuoto-ohjelma virheellisesti tulkitsee sanakirjassaan olevan sanan esiintymäksi tai sitten
    2b) käyttäjä on antanut hakusanan taivutusmuodossa.

Esimerkkinä ryhmästä 1) olkoon Alkula-sana, jota ei löydy sanakirjasta. Tallennusvaiheessa muodot Alkula, Alkulan, Alkulaa jne. jäävät tunnistamatta ja joutuvat tunnistamattomien sanojen hakemistoon. Sen sijaan Alkulakin, Alkulasta, Alkulana, Alkulankin jne. päätyvät perusmuotohakemistoon (perusmuotoina Alkulakki, Alkulasta, Alkulapsi, Alkulana, Alkulankki jne.) Hakuvaiheessa pitäisi löytää sekä perusmuoto- että tunnistamattomien sananmuotojen hakemistoon tallennetut hakemistosanat.

Ongelmakyselyistä ryhmään 1) kuuluvia, **aidosti tuntemattomia perusmuotoja** olivat tutkimusaineistossa hakusanat *muumi* (kysely 39), *Ihonen* (41), *Teräväinen* (42) ja *takinkääntäjä* (45; tuntematon oli yhdyssanan ensimmäinen osa, takki). Näitä sanoja ei sisältynyt Morfo-ohjelman tässä tutkimuksessa käytetyn version sanakirjaan.

Ryhmä 2a) käyttäytyy monilta osin samoin kuin ryhmä 1) edellä. Hakusana on tässäkin tapauksessa tuntematon perusmuoto. Perusmuoto-ohjelma ei kuitenkaan pidä hakusanaa tuntemattomana sanana, vaan tulkitsee sen toisen sanan taivutusmuodoksi. Jos hakusana on esimerkiksi Sulin, jota ei löydy sanakirjasta, se voidaan tulkita sulka-, sulaa- tai sula-sanan esiintymäksi.

Tulkintaohjelman sanakirjassa siis on sellainen perusmuoto, jonka jokin taivutusmuoto on identtinen hakusanan kanssa (sananmuotohomografi).

Ryhmän 2a) sananmuotoja tallennettaessa hakusanan perusmuoto ja mahdollisesti osa taivutusmuodoista tulkitaan väärin sananmuotohomografiensa esiintymiksi, joten ne tallennetaan väärän perusmuodon yhteyteen. Eihomografiset taivutusmuodot (esimerkiksi Sulinin, Sulinia) taas päätyvät tunnistamattomien sanojen hakemistoon. Ryhmään 2a) kuuluvia **sananmuotohomografeikseen tulkittuja perusmuotoja** olivat *Inga Sulin* (hakupyyntö 31), *Salmén* (käytännössä *Salmen,* ilman heittomerkkiä; kysely 40) ja *Sri Lanka* (kysely 44).

Ryhmässä 2b) hakusana oli perusmuoto-ohjelman sanakirjassa, mutta hakija antoi sen taivutusmuodossa. Testikyselyillä simuloitiin tilannetta, jossa hakija joko ei mieltäisi hakusanaa taivutusmuotoiseksi tai sitten perusmuotoohjelman käyttämä perusmuoto tuntuisi hakijan mielestä keinotekoiselta. Tällaisia olivat sanaliitot *Tarton rauha* (kysely 32), *Kansan Uutiset* (33), *Vuoden kylä* (34), *Yhtyneet Paperitehtaat* (35), *Seitsemän veljestä* (36), *Suomen Pankki* (37) ja *hallittu rakennemuutos* (38).

Tämäntyyppisen pulman syynä voi olla attributiivi, joka ei taivu pääsanansa mukaan, mutta ei ole perusmuodossakaan; suomen kielessä genetiiviattribuutti kuten *Suomen* Pankki on hyvin yleinen. Vaikka hakija tietäisikin, mikä on sanan perusmuoto, sen käyttäminen ei tuntuisi luontevalta. Sitä paitsi hakija ei voi käytännössä tietää, miten kukin yksittäinen sana on tallennusvaiheessa tulkittu (esimerkiksi hallittu -> hallita, yhtyneet -> yhtyä), koska se riippuu siitä, mitä sanoja perusmuoto-ohjelman sanakirjassa on. - Käytännössä hakujärjestelmässä pitäisikin olla oma toimintonsa, jolla hakija tarkistaa hakusanan tulkinnan: kun hakija syöttää tarkistettavan hakusanan, hakujärjestelmä palauttaa tiedon, miten kyseinen sana on tallennusvaiheessa tulkittu.

10.2 Automaattisesti katkaistujen hakusanojen tarkistus taivutusmuotohakemistosta haettaessa

10.2.1 Aidosti tuntematon tai sananmuotohomografikseen tulkittu perusmuoto

Jos hakusanan automaattinen katkaisu toteutetaan niin, että hakija syöttää hakusanan ja sen sanaluokan vartalo-ohjelmalle, tuntemattomat (ryhmä 1)

tai sananmuotohomografikseen tulkitut hakusanat (ryhmä 2a) eivät tuota erityisiä ongelmia. Kyselyhän toteutetaan tällöin samalla tavalla kuin vakiokyselyjenkin tapauksessa.

T2-tutkimusympäristön automaattisesti katkaistuilla muumi-, Ihonen- ja Teräväinen-sanoilla (ryhmä 1) saatiin tulokseksi samat dokumentit kuin hakijan itse katkaisemilla hakusanoilla. *Muumi*-kyselyn tulokseksi tuli 27 dokumenttia, joista oikeita 17 kappaletta ja lisäksi 5 dokumenttia, joissa muumi-sana oli yhdyssanan alkuosana. Vääriä dokumentteja oli 5; niissä esiintyi muumio tai sen johdoksia. *Ihonen*-kyselyn tuloksena saatiin 9 dokumenttia, joista 4 oikein, yksi oli sananmuotohomografi (iho-sanan muoto ihosta) ja 4 muuta sanaa, jotka alkoivat samalla tavalla kuin hakusana. *Teräväinen*-kyselyllä saatiin 17 dokumenttia, jotka kaikki olivat oikeita.

Samaten automaattisesti katkaistut Inga Sulin, Salmen ja Sri Lanka (ryhmä 2a) tuottivat samat dokumentit kuin hakijan katkaisemat hakusanat. *Inga Sulin* -kyselyllä saatiin 2 dokumenttia ja *Sri Lanka* -haussa 8 dokumenttia, kaikki oikeita. *Salmen*-kyselyllä saatiin 36 dokumenttia, joista oikeita oli 8, homografeja (esimerkiksi salmi-sanan genetiivejä) sisältäviä 14 kappaletta ja 14 sellaista dokumenttia, joissa oli hakusanan sisältävä yhdyssana (esimerkiksi Salmenhaara - ei siis haun kannalta relevantti).

Automaattisesti katkaistulla takinkääntäjä-sanalla saatiin vain osa niistä dokumenteista, jotka saatiin hakijan katkaisemilla hakusanoilla. Hakijan katkaisema hakusana näet oli lyhyempi ja palautti myös johdokset eli yhteensä 5 dokumenttia, joissa oikeita hakusanan muotoja oli kahdessa ja hakusanan johdoksia kolmessa dokumentissa. Automaattisen katkaisun tuloksena saatiin aluksi vain ne 2 dokumenttia, joissa varsinainen hakusana oli esiintynyt. Kun taivutusvartalo-ohjelmille syötettiin myös takinkääntäjä-hakusanan johdosperhe, löydettiin puuttuvatkin johdokset sekä näiden lisäksi yksi uusi johdos mukaan hakutuloksiin.

Mikäli hakusanan automaattinen katkaisu toteutetaan siten, että hakijan ei tarvitse itse antaa hakusanan sanaluokkaa, vaan se määritellään perusmuoto-ohjelman sanakirjan avulla, ryhmiin 1) ja 2a) kuuluvien hakusanojen käyttämisessä tulee vaikeuksia - näiden perusmuotojahan ei sanakirjassa ole. Tällöin niiden sanaluokka on määriteltävä Finstems- tai Hahmotin-ohjelmalle jollain muulla tavalla.

10.2.2 Taivutusmuodossa annettu hakusana

Kun taivutusmuotoiset hakusanat syötettiin taivutusvartalo-ohjelmille sellaisinaan (ja oikealla sanaluokalla varustettuina), tulosjoukko usein pieneni verrattuna hakijan katkaisemilla hakusanoilla saatuun tulosjoukkoon. Syynä oli se, että vartalo-ohjelmat joko tuottivat väärästä syötteestä vääriä vartaloita tai palauttivat syötetyn sanan muuttumattomana, joskin katkaisumerkillä varustettuna. Genetiiviattribuuttien (Tarton, Kansan, vuoden, Suomen) tapauksessa tosin ei haittaa, vaikka vartalo-ohjelma palauttaa taivutusmuotoisen hakusanan sellaisenaan, koska nämä sanat eivät taivu pääsanansa mukaan, vaan esiintyvät tekstissä juuri genetiivimuodossa.

Kun esimerkiksi *Kansan Uutiset* haettiin hakijan katkaisemilla hakusanoilla, saatiin 50 dokumenttia, joista 49 oikeita. Finstems-ohjelma katkaisi nämä taivutusmuotoiset hakusanat niin, että kyselyn tulokseksi saatiin vain 26 dokumenttia. Lähes puolet oikeista dokumenteista jäi siis löytymättä.

Toisaalta hakijan katkaisemissa hakusanoissa oli hyödynnetty genetiiviattribuutit tai muut vastaavat tapaukset, joissa sanaliiton alkuosa ei taivu (*Suomen* pankki), siten, että näissä tapauksissa sanaliiton alussa esiintyneen sanan loppuun ei ollut lisätty katkaisumerkkiä. Kun tällainen hakusana syötettiin vartalo-ohjelmalle, se lisäsi hakusanan loppuun katkaisumerkin, jolloin tulosjoukkoon ilmestyi ylimääräisiä dokumentteja, joissa hakusana oli esiintynyt yhdyssanan alkuosana. Automaattista katkaisua ei siis pitäisi tehdä hakujärjestelmässä pakolliseksi, vaan hakijalla pitäisi aina olla mahdollisuus katkaista hakusana itse tai jättää se kokonaan katkaisematta.

Esimerkiksi *Kansan Uutiset* -kyselyssä saatiin Hahmotin-ohjelman avulla haettaessa tulokseksi 64 dokumenttia eli 14 enemmän kuin hakijan katkaisemilla hakusanoilla haettaessa. Näissä 14 lisädokumentissa jompi kumpi hakusanoista oli esiintynyt yhdyssanan osana. Vaikka hakusanan katkaisu ja jokerimerkin käyttö laajensikin haun alaa ja lisäsi väärien dokumenttien määrää, on kuitenkin huomattava, että näitä lisädokumentteja ei olisi tullut lainkaan, jos BASIS-hakujärjestelmässä olisi ollut käytettävissä virkeoperaattoria tarkempi läheisyysoperaattori. Jos hakusanat olisi voitu rajata esiintymään vierekkäin, olisi sekä hakijan katkaisemilla että Hahmottimen katkaisemilla hakusanoilla saatu vain ne dokumentit, joissa todella esiintyi oikea sanaliitto.

Monikkomuotoisista hakusanoista Yhtyneet Paperitehtaat Hahmotin tuotti täysin oikeat vartalot. Kyselyllä *yhtyne** JA (*paperitehdas** TAI *paperitehtaa** TAI *paperitehtai** TAI *paperitehdaa** TAI *paperitehdai** TAI *paperitehdaks**) saatiin 93 dokumenttia, näistä oikeita eli sanaliiton sisältäviä dokumentteja 92. Yhdessä dokumentissa hakusanat esiintyivät samassa virkkeessä, mutta eivät muodostaneet haettua sanaliittoa. Finstems puolestaan palautti syötetyt monikkomuodot sellaisinaan. Sen tuottamilla vartaloilla *yhtyneet** JA *paperitehtaat** saatiin tulokseksi vain 54 dokumenttia, tosin kaikki oikeita. Hakijan katkaisemilla hakusanoilla toteutettu kysely (T1) *yhtyne** JA *paperiteh** tuotti kaikkiaan 96 dokumenttia. Vain hakijan katkaisemilla hakusanoilla löydettiin ne kolme dokumenttia, joissa jälkimmäinen hakusana esiintyi lyhenteenä Paperiteht.

Yhtyneet on paitsi monikkomuoto, myös verbin partisiippimuoto. Tämä ja toinen esimerkkikyselyissä esiintynyt partisiippimuoto, hallittu, eivät tuottaneet Hahmotin-ohjelmalle ongelmia, vaan se muodosti niistä sopivat taivutusvartalot. Finstems ei taivuttanut yhtyneet-sananmuotoa lainkaan, mutta hallittu-sanan se taivutti oikein. Numeraalin seitsemän Hahmotin taivutti oikein, kun taas Finstems-ohjelman testiversio ei osannut taivuttaa sitä ollenkaan, vaan palautti sanan sellaisenaan. Veljestä-taivutusmuoto oli jo niin hankala, ettei kumpikaan vartalo-ohjelmista tuottanut siitä oikeita vartaloita.

Tutkimusympäristössä T2 hakijan siis on oltava tarkkana ja annettava hakusana oikeassa eli perusmuodossa, jotta myös vartalo-ohjelmat tuottavat oikeat vartalot. Erityisesti Finstems edellytti, että hakusana annettiin tarkalleen määrittelyjen mukaisesti. Tämä ongelma voidaan kiertää siten, että ennen vartaloiden tuottamista tarkistetaan, onko hakusana perusmuodossa. Ensin hakusana syötetään perusmuoto-ohjelmalle, joka tarvittaessa perusmuotoistaa hakusanan. Sitten perusmuoto-ohjelman antama perusmuoto ja sanaluokka syötetään taivutusvartalo-ohjelmalle. Kun tällaista tarkistusta kokeiltiin manuaalisesti tutkimusympäristössä T2, ei dokumentteja enää karsiutunut puutteellisten vartaloiden vuoksi. Itse asiassa tulosjoukot sisälsivät dokumentteja enemmän kuin hakijan katkaisemilla hakusanoilla haettaessa. Perusmuotoistetuista sanoista tuotetuilla vartaloilla nimittäin saatiin tulokseksi hakusanan kaikki taivutusmuodot - myös ne, jotka hakijan katkaisemilla hakusanoilla haettaessa oli tarkoituksella jätetty pois.

Kun Finstems-ohjelmalle syötettiin oikeat perusmuodot *yhtyä* JA *paperitehdas*, kysely muuttui muotoon (*yhty** TAI *yhdy**) JA (*paperitehdas** TAI

*paperitehtai** TAI *paperitehtaa** TAI *paperitehdaks**), jolla saatiin 93 dokumenttia eli samat kuin Hahmotin-ohjelman tuottamilla vartaloilla. Kun *Kansan Uutiset* -kyselyn hakusanat syötettiin Hahmotin- ja Finstems-ohjelmille perusmuotoisina, taivutusvartaloilla tehdyn kyselyn tuloksena saatiin 92 dokumenttia, joista oikeita 49. Ylimääräisissä 42 dokumentissa hakusanat olivat esiintyneet yhdyssanan osina tai jossain muussa kuin halutuissa taivutusmuodoissa. Kaikki ylimääräiset dokumentit olisivat kuitenkin karsiutuneet pois, jos BASIS-hakujärjestelmässä olisi voinut käyttää tarkempaa läheisyysoperaattoria.

10.2.3 Yhteenveto

Jos automaattisen katkaisun oikeellisuus halutaan varmistaa, on taivutusmuotoiset sanat ensin palautettava Morfo- tai Twol-ohjelman avulla perusmuotoon ja vasta tämä perusmuoto syötettävä vartalo-ohjelmalle. Toisaalta taipumattomia sananmuotoja, kuten genetiiviattribuutteja, ei pitäisi katkaista lainkaan. Sanaliittoja haettaessa on olennaisen tärkeää, että hakujärjestelmässä on riittävän tarkka läheisyysoperaattori, joilla hakusanat voidaan määritellä esiintyväksi esimerkiksi välittömästi peräkkäin tietyssä järjestyksessä.

Jos hakusana voidaan tunnistaa jonkin sanakirjassa olevan sanan taivutusmuodoksi, ei sen jatkokäsittelyä pitäisi kuitenkaan automatisoida liikaa. Ei siis niin, että tällainen hakusana heti perusmuotoistetaan ja sitten syötetään perusmuotoistettu hakusana vartalo-ohjelmalle, jonka tuottamat vartalot lisätään kyselyyn. Kaikki taivutusmuotoisiksi tulkittavissa olevat sanat kun eivät välttämättä sellaisia ole - kyseessähän voi olla aito perusmuoto, joka vain tulkitaan homografikseen. Tällainen sananmuoto on käsitelteltävä 10.2.1-luvussa kuvatulla tavalla.

10.3 Syötteen tarkistaminen perusmuotohakemistoista haettaessa

10.3.1 Aidosti tuntematon perusmuoto

Tutkimusympäristöissä T4 (perusmuotohakemisto) ja T5 (ositettu perusmuotohakemisto) ilman tulkintaa jäävät sananmuodot tallennetaan tunnistamattomien sananmuotojen hakemistoon. Ongelmana on, että vaikka hakusanan perusmuoto itse puuttuisikin sanakirjasta, osa sen taivutusmuodoista voi olla identtisiä jonkin sanakirjassa olevan sanan perus- tai taivutusmuodon kanssa. Tallennusvaiheessa tällaiset sananmuotohomografit tulkitaan

yksinomaan tämän sanakirjassa olevan sanan esiintymiksi, jolloin ne tallennetaan perusmuotohakemistoon, mutta väärän perusmuodon yhteyteen (vrt. Alkula, Alkulaa -> tuntemattomia; Alkulakin -> Alkulakki).

Koska sanan eri esiintymät voivat joutua eri hakemistoihin, on tämäntyyppisiä ongelmasanoja haettaessa käytävä läpi sekä varsinainen (ositettu) perusmuotohakemisto että tunnistamattomien sananmuotojen hakemisto. FULLTEXT-projektin tutkimusympäristöissä tämä toteutettiin syöttämällä tunnistamatta jääneet hakusanat ja niiden sanaluokka taivutusvartaloita tuottavalle Finstems-ohjelmalle. Sanaluokan määritteli hakija eli tässä tapauksessa tutkija itse. Finstems-ohjelman tuottamilla hakusanoilla haettiin molemmista hakemistoista (vartalolla Alkula* löydettäisiin yllämainitut esimerkkimuodot). Vastaavaa korjaushakua ei enää tehty Hahmotin-ohjelmalla, koska tässä tapauksessa sen ja Finstemsin vertailu ei olisi tuottanut projektin kannalta uutta tietoa.

Tämänkaltainen hakusanan automaattiseen katkaisuun perustuva ratkaisu oli mahdollinen, koska yhdyssanat oli tallennettu hakemistoon kokonaisina - jos ne olisi tallennettu erikseen, esimerkiksi muodossa alku- ja -lakki, olisi tarvittu mutkikkaampi korjausmenetelmä. Tämän menetelmän ongelmana on, että hakusanan sanaluokka pitää kysyä hakijalta, mikä edellyttää hakijalta jonkinlaista kieliopin tuntemusta eikä siten ole optimaalinen ratkaisu. Tuotantokäytössä olevaan hakujärjestelmään pitäisikin rakentaa käyttöliittymä, joka pystyisi avustamaan hakijaa sanaluokan määrittelyssä mahdollisimman pitkälle.

Tutkimusympäristöissä T4 ja T5 saatiin *muumi*- ja *Teräväinen*-kyselyissä sama tulos kuin hakijan katkaisemia hakusanoja käytettäessä. Sen sijaan *takinkääntäjä*-kyselyn tuloksena saatiin vähemmän dokumentteja, koska johdokset jäivät pois. Kun johdoksetkin lisättiin kyselyyn, saatiin tämän täsmennyksen avulla tulokseksi samat dokumentit kuin hakusanat automaattisesti katkaistaessa. Näin ollen näiden kaikkien kolmen kyselyn tulokseksi saatiin täsmälleen samat dokumentit kuin tutkimusympäristössä T2 (luku 10.2.1).

Ihonen-kyselyn tulokseksi saatiin perusmuotohakemistoista 7 dokumenttia, joista 4 oikein. Vastaava kysely T1- ja T2-tutkimusympäristöistä tuotti kaikkiaan 9 dokumenttia, joista oikeita dokumentteja oli saman verran eli 4. Loput dokumentit sisälsivät sanoja, jotka sattumalta alkoivat samalla tavalla

kuin hakusanojen alkuosa. Perusmuotohakemistoista haettaessa tulosjoukosta jäivät pois dokumentit, joissa esiintyneet taivutusmuodot (ihostaan, ihosta) oli tulkittu iho-sanan esiintymiksi. Korjauskyselyssä hakusanoina olivat ihonen- ja ihos-vartalot, jotka eivät palauta hakemistosta niitä lyhyempää perusmuotoa iho. Tosin kummassakin poisjääneessä dokumentissa todella oli iho-yleisnimen eikä Ihonen-erisnimen taivutusmuoto, joten tässä nimenomaisessa tapauksessa ei menetetty oikeita dokumentteja.

Vaikka *Ihonen*-ongelmakyselyssä ei sananmuotohomografien väärän tulkinnan vuoksi hävinnytkään dokumentteja, se ei kuitenkaan kumoa sitä seikkaa, että kaikkia tämäntyyppisiä tulkintavirheitä ei voi korjauskyselyilläkään paikata. Erityisesti yhdyssanat ovat hankalia. Vaikka yhdyssanaksi tulkittavasta sananmuodosta tallennetaan hakemistoon koko yhdyssana ja väärä tulkinta on periaatteessa jäljitettävissä hakemalla hakusanan vartaloilla, ei esimerkiksi keskusohjain-sanan eri vartaloilla löydetä hakemistosta perusmuotoa keskusohjatimeä, joka on saatu perusmuotoistamalla sananmuoto keskusohjaimen.

Mikäli ehdottomasti halutaan löytää kaikki mahdolliset dokumentit, joissa tietty (ongelma)sana on esiintynyt, se on periaatteessa mahdollista. Koskenniemi (1985c) on esittänyt ratkaisun, jossa ongelmasana (kuten Ihonen tai keskusohjain) sijoitetaan tilapäisesti kaksitasomallin sanakirjaan. Tämän jälkeen tästä perusmuodosta aletaan **tuottaa** taivutusmuotoja eli sitä sovelletaan päinvastaiseen suuntaan kuin perusmuotoistettaessa. Samalla kun uusia taivutusmuotoja tuotetaan, tuotetut muodot analysoidaan kaksitasomallin normaalilla, perusmuotoistavalla versiolla. Tämän analyysin avulla selvitetään, voidaanko käsiteltävänä oleva sananmuoto tulkita joidenkin muiden, sanakirjassa jo olevien sanojen taivutusmuodoksi. Lopputuloksena olisi joko tieto, ettei sananmuotohomografeja löydy, tai sitten luettelo sanoista, joiden jokin taivutusmuoto on ongelmasanan sananmuotohomografi. Jälkimmäisessä tapauksessa järjestelmä sitten käy läpi dokumentit, joissa esiintyy jokin luettelossa olleista sanoista, löytääkseen ne, joissa hakusana itse on esiintynyt. Menetelmän on yksityiskohtaisemmin kuvannut Hjorth (1987, s. 67 - 69.)

Edellä kuvattua sananmuotojen tuottamiseen perustuvaa tarkistusta ei FULLTEXT-projektin aikana ollut käytössä missään hakujärjestelmässä. Suomen kielelle toteutettuna tällainen tuottamisprosessi olisi raskas, koska tuotettavia sananmuotoja on tuhansia. Sen sijaan sellaisiin kieliin, jossa sa-

nat taipuvat vähemmän kuin suomessa, menetelmä soveltuisi paremmin. Esimerkiksi englannin mouse-sanalle tarvitsee tuottaa vain neljä sananmuotoa (mouse, mouse's, mice, mice's), monille muille sanoille vielä vähemmän. (Arppe 1996).

Yksi kiinnostavaa tutkimuksen aihe olisi selvittää, onko Koskenniemen malli ainoa kattava keino löytää väärintulkitut sananmuodot, vai voidaanko sen kanssa suunnilleen samaan lopputulokseen päästä myös heuristisin keinoin eli laatimalla tarkistusmenetelmä, joka kaikkien mahdollisten vaihtoehtojen sijasta tutkisi vain todennäköisimmät ongelmasanat. Laalo (1990, s. 59 - 60) mainitsee, että toisen nominin jonkin muun sijan kuin partitiivin, genetiivin tai essiivin kanssa identtisiä perusmuotoja on niukasti. On vain vähän sellaisia nomineja, kuten postilla, prinsessa ja kompleksi, joiden vartalo loppuu samanlaiseen merkkijonoon kuin jokin muu sijapääte. Vieläkin vähemmän on sellaisia taivutusmuotoja, joiden kanssa nämä erikoiset perusmuodot voisivat langeta yhteen - posti-sanan adessiivi sattuu olemaan tällainen harvinaisuus.

Hakujärjestelmään voitaisiin esimerkiksi rakentaa päättelysäännöt, että kun hakusana on perusmuoto-ohjelmalle tuntematon ja sen sanaluokka on selvitetty (esimerkiksi kysymällä hakijalta), näistä hakusanoista muiden tarkistusten lisäksi tuotettaisiin perusmuoto-ohjelmalla tietyt potentiaaliset ongelmamuodot. Nomineista muodostettaisiin ainakin partitiivi-, genetiivi- ja essiivimuodot, jotka sitten taas analysoitaisiin perusmuoto-ohjelmalla. Tällaisessa tarkistuksessa havaittaisiin Ihonen-nimen partitiivimuodon Ihosta olevan identtinen iho-sanan elatiivin kanssa. Samaten keskusohjain-sanan genetiivimuodon keskusohjaimen avulla voitaisiin hakemistosta jäljittää väärintulkinnan tuloksena syntynyt keskusohjatimeä-perusmuoto.

10.3.2 Sananmuotohomografikseen tulkittu perusmuoto

Sulin- ja *Salmen*-hakusanoja ei löytynyt sellaisenaan perusmuotohakemistosta, mutta ne voitiin tulkita sanakirjassa olevien sanojen taivutusmuodoiksi. Virhetulkinnat korjattiin menetelmällä, jossa hakemistoista etsittiin sekä väärin tulkitut että tunnistamatta jääneet sananmuodot. Ensin perusmuotohakemistosta poimittiin kaikki ne perusmuodot, joiden taivutusmuodoksi hakusanan perusmuoto oli voitu tulkita (Sulin -> sulka, sulaa, sula; Salmen -> salmi, Salme). Tämän lisäksi hakusanat ja niiden sanaluokat syötettiin Finstems-ohjelmalle, jonka tuottamilla taivutusvartaloilla haettiin molem-

mista hakemistoista (tämä vaihe oli sama kuin edellisessä 10.3.1-luvussa selostettu).

Kun hakusanana oli pelkkä *Sulin*, tulosjoukko sisälsi 171 dokumenttia (eli sula-, sulaa- ja sulka-sanojen esiintymät perusmuotohakemistossa sekä taivutusvartaloiden tuottamat osumat kummastakin hakemistosta), kun tämä sana hakijan katkaisemana tuotti taivutusmuotohakemistosta vain 15 dokumenttia. Vastaavasti *salmen*-hakusana tuotti 166 dokumenttia (joista 149 kappaletta salmi-sanan tuottamaa osumaa), kun hakijan katkaisemilla hakusanoilla saatiin taivutusmuotohakemistosta 36 dokumenttia. *Inga Sulin*-kyselyn lopputulokseksi saatiin samat kaksi dokumenttia kuin hakijan katkaisemilla hakusanoilla, koska Inga rajasi tehokkaasti pois väärät sulka-yms. sanojen esiintymät. Sen sijaan *Salmen*-kyselyssä vastaavaa karsijaa ei käytetty (vaikka *Leif* olisi mitä todennäköisimmin tuottanut vastaavan tuloksen).

Edelläkuvatun korjausmenetelmän ongelmat ovat vartalohaun osalta samat kuin edellisessä luvussa kuvatuissa esimerkeissä. Lisäksi tulee väärintulkittujen sananmuotohomografien aiheuttama epätarkkuus: vaikka vain yksi tietyn hakemistosanan sananmuodoista olisi hakusanan sananmuotohomografi, tulosjoukkoon päätyvät nekin dokumentit, joissa esiintyy tämän hakemistosanan muita taivutusmuotoja. Perusmuotohakemistostahan ei löydy yksittäisiä sananmuotoja, vaan vain niiden yhteinen perusmuoto. Tässä suhteessa perusmuotohakemistoa käytettäessä on enemmän tarkkuusongelmia kuin taivutusmuotohakemistoa käytettäessä.

Tarkkuusongelman korjaamiseksi tehtiin vielä yksi lisätarkistus: alkuperäinen hakusana syötettiin Finstems-ohjelmalle, jonka tuottamilla vartaloilla tehtiin **peräkkäishaku** itse dokumenttien teksteistä. Alkuperäisen hakusanan todelliset esiintymät siis haettiin dokumenteista, jotka edellisessä vaiheessa oli saatu kyselyn tuloksena. Koska sanat alkuperäistekstissä olivat taivutusmuodossaan, vartalot palauttivat vain hakusanan taivutusmuotohomografit eivätkä enää muita hakemistosanan taivutusmuotoja.

Tällaisen karsinnan tuloksena *Sulin*- ja *Salmen*-hakusanoilla saadut lopulliset tulosjoukot sisälsivät samat dokumentit kuin hakijan katkaisemilla hakusanoilla taivutusmuotohakemistosta haettaessa (eli Sulin-sanalla 15 ja Salmen-sanalla 36 dokumenttia). Hakutuloksen tarkkuus saatiin siis varsin yksinkertaisella menetelmällä nostettua vastaavalle tasolle kuin hakijan kat-

kaisemia hakusanoja käytettäessä. Menetelmän haittapuolena on, että jos tulosjoukot ovat suuria ja dokumentit pitkiä, peräkkäishaku kestää kauan. Muuten peräkkäishaku ei tuota hakijalle lisävaivaa, onhan esimerkiksi hakusanojen sanaluokka pitänyt määritellä jo tarkistusten aiemmassa vaiheessa.

Erityisiä ongelmia perusmuotohakemistoissa tuotti *Sri Lanka*-kysely, jonka saanti oli huonompi kuin käytettäessä hakijan katkaisemia hakusanoja taivutusmuotohakemistossa. Tallennusvaiheessa Lanka-, Lankaa- ja Lankaan-muodot oli liitetty lanka-yleisnimeen, kun taas Lankan, Lankassa jne. joutuivat tulkitsemattomina tunnistamattomien sananmuotojen hakemistoon. Koska lanka-sana hakuvaiheessa löytyi suoraan perusmuotohakemistosta, homografiongelmaa ei havaittu ja eikä korjauskyselyä tunnistamattomien sananmuotojen hakemistosta tehty. Kun taivutusmuotohakemistosta löydettiin 8 dokumenttia, perusmuotohakemistoista löydettiin vain 2 dokumenttia. Loput 6 jäivät löytymättä, koska niissä olevat Lanka-sanan esiintymät olisi löydetty vain tunnistamattomien sananmuotojen hakemistosta.

10.3.3 Taivutusmuodossa annettu hakusana

Kun perusmuotohakemistoista piti hakea taivutusmuodossa annetuilla hakusanoilla, korjauskysely aloitettiin perusmuotoistamalla hakusanat. Näin simuloitiin hakujärjestelmää, jossa perusmuoto-ohjelma on käytettävissä sekä tallennus- että hakuvaiheessa: hakijan antamat hakusanat palautetaan automaattisesti samaan muotoon kuin tallennusvaiheessa ja sitten haetaan näillä perusmuodoilla. Näin ongelmakyselyjen hakusanat muuntuivat seuraavanlaisiksi: *tartto* JA *rauha*, *kansa* JA *uutinen*, *vuosi* JA *kylä*, *yhtyä* JA *paperitehdas*, *seitsemän* JA *veli*, *suomi* JA *pankki* sekä *hallita* JA *rakennemuutos* (ja lisäksi niin, että hakusanat oli kytketty toisiinsa myös virkeoperaattorilla JA-operaattorin sijasta).

Kun alkuperäiset hakusanat korvattiin kyselyissä perusmuodoilla, näin saatujen tulosjoukkojen tarkkuusarvot yleensä jäivät alemmiksi kuin hakijan katkaisemilla hakusanoilla taivutusmuotohakemistosta saaduissa tulosjoukoissa. Hakuahan ei voinut rajata vain tiettyihin taivutusmuotoihin.

Perusmuotohakemistosta tehtyjen kyselyjen tarkkuutta kokeiltiin parantaa siten, että halutut taivutusmuodot poimittiin peräkkäishaun avulla esiin tulokseksi saatujen dokumenttien varsinaisesta tekstistä (kuten edellisessä luvussa 10.3.2 on kuvattu). Alkuperäiset taivutusmuotoiset hakusanat syötet-

tiin Finstems-ohjelmalle, jonka tuottamilla vartaloilla tehtiin peräkkäishaku. Näin toimien tulokseksi saatiin yleensä samat dokumentit kuin tutkimusympäristössä T2, jossa Finstems-ohjelman katkaisemilla hakusanoilla haettiin taivutusmuotohakemistosta (luku 10.2.2).

Käytännössä tämä ratkaisu ei kuitenkaan toimi, sillä jos hakusanat eivät ole perusmuodossa, ne tuottavat vääriä vartaloita. Tällöin taas osa oikeista dokumenteista jää löytymättä eli saanti kärsii. Toisaalta kaikkia ongelmia ei tarvitsekaan yrittää ratkaista lingvistisin keinoin - monet epätarkkuudet voidaan korjata, jos hakujärjestelmässä on käytössä riittävän tarkka läheisyysoperaattori.

Perusmuotohakemistoista saadut tulosjoukot erosivat taivutusmuotohakemistosta saaduista tulosjoukoista eniten *Vuoden kylä* -kyselyssä (hakupyyntö 34). Kun kysely tehtiin taivutusmuotohakemistosta joko hakijan katkaisemin tai automaattisesti katkaistuin hakusanoin, väärät osumat johtuivat siitä, että hakusanat olivat esiintyneet yhdyssanan osina. Perusmuotohakemiston väärät osumat taas saatiin siksi, että niissä esiintyi sananmuoto vuonna, joka oli perusmuotoistettu vuosi-sanaksi. Sekä perusmuoto- että taivutusmuotohakemistosta väärät osumat olisivat jääneet tulematta, jos hakujärjestelmässä olisi ollut käytettävissä virkeoperaattoria tarkempi läheisyysoperaattori.

Perusmuotoisista hakemistoista haettaessa *Tarton rauha*-kyselyn tulokseksi saatiin 5 dokumenttia. Taivutusmuotohakemistosta haettaessa saatiin tulokseksi 9 dokumenttia eli samat 5 dokumenttia kuin edellä ja lisäksi 4, joissa rauha oli yhdyssanan alkuosana: Tarton *rauhansopimus*.

Kokeiltu korjausmenetelmä kaipaisi jatkokehittelyä: koska hakusanat eivät olleet perusmuodossa, Finstems ei luonnollisestikaan tuottanut niistä oikeita vartaloita. Lisäongelmana olivat ne perusmuodot, kuten seitsemän, joita tutkimuksessa käytetty Finstems-ohjelman versio ei osannut taivuttaa ollenkaan. Korjauksia ei siis voi toteuttaa suoraviivaisesti vain syöttämällä hakusanat normaalioletuksin toimivalle vartalo-ohjelmalle. Toisaalta tuloksen tarkkuutta voidaan helposti parantaa muillakin kuin lingvistisillä keinoilla: tarkkuus olisi ollut selvästi parempi, jos sanaliittojen hakusanat olisi voitu kytkeä toisiinsa virkeoperaattoria tarkemmalla läheisyysoperaattorilla.

10.4 Syötteen tarkistaminen kaksoishakemistoista haettaessa

Kaksoishakemistosta eli tutkimusympäristöstä T6 haettaessa periaatteena oli, että kun perusmuoto-ohjelma pystyi tunnistamaan hakusanan, haettiin samalla tavalla kuin normaalisti tutkimusympäristössä T4 tai T5, siis perusmuotoisilla hakusanoilla perusmuotohakemistosta. Ongelmatapauksissa eli kun hakusanat puuttuivat perusmuoto-ohjelman sanakirjasta, kyselyt suoritettiinkin taivutusmuotohakemistosta.

Ongelmakyselyjen tulokset riippuvat sitten siitä, haetaanko hakijan katkaisemilla hakusanoilla, jolloin tulokseksi saadaan samat dokumentit kuin tutkimusympäristöstä T1. Jos taas hakusanat katkaistaan vartalo-ohjelmalla, saadaan samat tulokset kuin tutkimusympäristöstä T2.

Varsinaisesti vain *Sri Lanka*-kyselyn tulokset poikkesivat hakijan katkaisemilla hakusanoilla toteutettujen kyselyjen (T1) tuloksista. Koska lanka-sana oli perusmuoto-ohjelman sanakirjassa, Lanka-nimi tulkittiin homografikseen, jolloin perusmuotoisen hakusanan käytössä ei havaittu ongelmia. Siksi kyselyn tulokseksi saatiin vain perusmuotohakemiston kautta löytyneet 2 dokumenttia. Loput 6 jäivät löytymättä, koska ne olisi löydetty vain hakemalla myös taivutusmuotohakemistosta. Toinen poikkeama oli *Tarton rauha*-kysely, jonka tulokseksi saatiin perusmuodoilla hakien vain 5 dokumenttia. Löytymättä jäivät ne 4 dokumenttia, joissa rauha oli esiintynyt rauhansopimus-yhdyssanan osana (luku 10.3.3.); nekin tosin olisivat löytyneet perusmuotohakemistosta, jos myös yhdyssanat olisi otettu kyselyyn mukaan.

11 TULOSTEN TARKASTELU

11.1 KOEJÄRJESTELYT

FULLTEXT-projektin testijärjestelyiden suunnittelu oli monessa suhteessa pioneerityötä. Vaikka erityisesti angloamerikkalaiset tutkijat ovat tutkineet, miten hakusanojen pääteainesten karsinta vaikuttaa hakutulokseen, ja vaikka myös haku- ja hakemistosanojen muuntamista jonkinlaiseen perus- tai normaalimuotoon on tutkittu (luku 4.3.), aivan tämänkaltaista tutkimusta ei ole aiemmin tehty.

Relevanssiarvioiden tekemisessä otettiin mallia Tenopirin ja Ron (1990) tutkimuksesta, jossa käytettiin kolmen asiantuntijan raatia. Eri kyselytyyppien laatimisessa, kyselyjen laatimisessa ja merkitsevyystesteissä mallina olivat Kristensenin (1992; 1993) tutkimuksessa sovelletut menetelmät. Esikuvina olivat kuitenkin vain toteutustavat ja menetelmät (kuten kyselyn vaiheittainen laajentaminen erityyppisiä hakusanoja lisäämällä), eivät tutkimushypoteesit, hakemistojen toteutustavat, tutkimusympäristöt tai kyselytyypit - nämä tutkija suunnitteli FULLTEXT-projektia varten itse.

Projektin tutkimustietokanta sisälsi 23 244 dokumenttia ja testikyselyinä oli 26 vakiokyselyä ja 15 ongelmakyselyä. Tutkimusaineisto oli siis riittävän suuri ja siten edustava verrattuna alan aiempiin tutkimuksiin. Esimerkiksi 1980- ja 1990-luvuilla usein käytetyssä CACM-kokoelmassa oli 3 204 dokumenttia ja 52 kyselyä, joille on koottu valmiit relevanssiarviot; INSPEC-kokoelmassa 12 684 dokumenttia ja 77 kyselyä, sekä MED-kokoelmassa 1 033 dokumenttia ja 30 kyselyä relevanssiarvioineen (Salton 1989, s. 358).

Nykyään informaatiotutkijoiden pyrkimyksenä on käyttää testauksissa realistisen kokoisia, mielellään satoja tuhansia tai miljoonia dokumentteja sisältäviä tutkimustietokantoja. Esimerkiksi TREC-1-kokoelma sisälsi erityyppisiä dokumentteja yli 700 000 kappaletta ja hakupyyntöjä valmiine relevanssiarvioineen oli 50 kappaletta (Harman 1993). Vuonna 1997 TREC-6-konferenssissa erityyppisiä dokumentteja oli jo useita miljoonia. Esimerkiksi uutisia oli lähes 850 000 kappaletta, joista pelkästään Financial Timesin uutisia yli 200 000 kappaletta, ja tiedonhakua simuloivia hakupyyntöjä (topic) oli 350 kappaletta (Sparck Jones 2000). - Erikoisemmilla kielija tutkimusalueilla joudutaan kuitenkin nykyisinkin tyytymään pienempiin

aineistoihin: vuonna 1998 TREC-7-konferenssissa puhedokumenttien haussa oli aineistona 2 866 dokumenttia (radiouutista) ja hakupyyntöjä oli 23; seuraavana vuonna TREC-8-konferenssin aineisto oli kasvanut jo 21 754 dokumenttiin ja 50 kyselyyn (Johnson et al. 1999).

FULLTEXT-projekti lähti käyntiin kiinnostuksesta selvittää perusmuotoistamisen vaikutus tiedonhaun tuloksiin. Yksi projektin suunnitteluvaiheen oivallus oli, että hakijan itse katkaisemien hakusanojen vertailu pelkästään hakusanojen perusmuotojen kanssa olisi liian suppeaa, koska johdosten vaikutus jäisi tällöin huomiotta. Eri vaihtoehtojen huomioonottamiseksi oli siis suunniteltava useita kyselytyyppejä. Tutkimusta varten laaditut kyselytyypit vaihtelivat suppeasta, vain hakusanojen perusmuodot sisältävästä peruskyselystä aina laajimpaan osien yhdistelmäkyselyyn, joka sisälsi alkuperäisten hakusanojen lisäksi myös hakusanojen johdosperheen sekä yhdyssanat, joissa hakusana tai sen johdosperheen jäsen oli osana. Erilaisia kyselytyyppejä oli kaikkiaan kahdeksan. Vastaavaa jaottelua, joka ottaisi erilaiset johdos- ja yhdyssanavaihtoehdot huomioon, ei ole tehty ulkomaisissa tutkimuksissa. Vaikka aiemmissa suomalaisissa tutkimuksissa on pohdittu yhdyssanojen vaikutusta hakutulokseen, ei niissä juurikaan ollut kiinnitetty huomiota johdoksiin (paitsi Kristensen & Järvelin 1990).

Tiettävästi missään aiemmassa tutkimuksessa ei ole erityisen perusteellisesti pohdittu niitä tekijöitä, joita yhdyssanojen osittamisessa ja osien tallentamisessa hakemistoon tulisi ottaa huomioon. Tällainen analyysi tehtiin FULLTEXT-projektissa (luku 7) ja analyysin perusteella päätettiin, mitkä tutkimusympäristöt useista mahdollisista vaihtoehdoista projektissa kannattaisi toteuttaa. FULLTEXT-projektin kokemuksia on sittemmin hyödynnetty muun muassa Tampereen yliopiston tiedonhaun tutkimuslaboratoriossa INQUERY-hakujärjestelmän perusmuotohakemiston rakentamisessa (Keskustalo 1994).

Tutkimusaineisto jaettiin 26 kyselyn perusjoukon lisäksi vielä kahteen osajoukkoon, johdososajoukkoon ja yhdyssanaosajoukkoon, joissa eri kyselytyyppien vaikutusta hakujen tuloksiin tarkasteltiin yksityiskohtaisemmin. Lisäksi tehtiin jako vakio- ja ongelmakyselyihin; jälkimmäisille suunniteltiin ja kokeiltiin menetelmiä, joilla voidaan korjata perusmuoto-ohjelmien tekemät väärintulkinnat.

Relevanssiarviot toteutettiin Tenopirin ja Ron (1990) selostaman tutkimuksen mukaisesti, ts. relevanssin arvioi kolme asiantuntijaa ja enemmistön mielipide ratkaisi dokumentin relevanssiluokituksen. Lisäksi tutkija laati relevanssiarvioiden tueksi lyhyen kehyskertomuksen, jonka avulla dokumenttien relevanssi voitiin arvioida täsmällisemmin kuin pelkän lyhyen hakupyynnön perusteella (Borlund & Ingwersen 1997). Tällaista ratkaisua ei ennen tätä tutkimusta tiettävästi ollut sovellettu suomalaisissa tutkimuksissa; vastaava menetelmä oli käytössä myös Sormusen (1994) tutkimuksessa, joka toteutettiin FULLTEXT-projektin jälkeen.

Kehyskertomuksen käyttö ei ole ainutlaatuinen ratkaisu, vaan sen tyyppisiä hakupyyntöä laajempia kuvauksia on otettu käyttöön muun muassa TREC-hankkeessa. TREC-hankkeessa kutakin hakupyyntöä varten on laadittu varsin yksityiskohtaiset kuvaukset siitä, millaiset ominaisuudet dokumentilla on oltava, jotta se katsotaan hakupyynnön kannalta relevantiksi (Harman 1993).

FULLTEXT-projektin testausjärjestelyjen osalta voidaan vielä todeta, että testiympäristön rakentaminen, kyselyjen suorittaminen ja hakutulosten relevanssiarviointi on erittäin työläs ja paljon käsityötä vaativa tutkimustapa. Koska tutkimuksessa oli paljon erilaisia muuttujia, tulosten analysointiin ja merkitsevyystestien laskemiseen kului erittäin paljon aikaa. Vastaavan tutkimuksen toteuttaminen toisentyyppisillä kyselyillä tai toisella aineistolla on käytännössä pakko tehdä yleisessä, standardoidussa tutkimusympäristössä, kuten Tampereen yliopiston informaatiotutkimuksen laitoksen tiedonhakulaboratoriossa; muuten tällaisessa empiirisessä laboratoriotutkimuksessa jo pelkästään testausympäristön rakentaminen vaatii suhteettoman suuren työmäärän. Tätä tutkimusta aloitettaessa ei tällaista yleistä tutkimusympäristöä kuitenkaan ollut käytettävissä, sillä Tampereen tiedonhakulaboratoriokin rakennettiin vasta FULLTEXT-projektin jälkeen (Keskustalo 1994).

11.2 HYPOTEESIEN TOTEUTUMINEN

Seuraavassa tarkastellaan tutkimushypoteesien esittämien väittämien paikkansapitävyyttä.

Hypoteesi 1

> Kun tekstin sananmuodot palautetaan perusmuotoon ennen kuin ne tallennetaan hakemistoon, perusmuotohakemistoon tulee vähemmän

erilaisia merkkijonoja kuin vastaavasta tekstistä tuotettuun taivutusmuotohakemistoon. Koska hakemiston merkkijonojen määrä vähenee, perusmuotohakemisto vie kilotavuissa mitaten vähemmän muistitilaa kuin samasta tekstistä tuotettu taivutusmuotohakemisto. (Siis: Hn < H1, missä n = 2 tai 3.)

Ensimmäisen tutkimushypoteesin väite osoittautui paikkansapitäväksi (luku 8): Suomenkielisessä tekstissä yksittäisestä sanasta esiintyy paljon erilaisia sananmuotoja. Siten taivutusmuotohakemistoon joudutaan tallentamaan paljon erilaisia merkkijonoja, mikä vaatii paljon muistitilaa. Kun eri sananmuodot tiivistetään hakemistossa yhdeksi perusmuodoksi, muistitilaa säästyy (H2 < H1). Vielä silloinkin, kun perusmuotohakemisto sisälsi myös yhdyssanojen osat ja kaikki niiden yhdistelmät, se oli pienempi kuin taivutusmuotohakemisto (H3 < H1).

Hakemisto siis kutistuu, kun käsitellään suomenkielisiä tekstejä. Suomen kielessä on kuitenkin suhteessa vähemmän monitulkintaisia sananmuotoja kuin esimerkiksi ruotsissa ja englannissa (Karlsson 1994, s. 80). Homografisista sananmuodoista on tallennettava hakemistoon kaikki mahdolliset tulkinnat, joten suomesta poikkeavissa kielissä homografien suurempi osuus voi osaltaan kumota perusmuotoistamisen etuja. Monitulkintaisten ilmausten yksiselitteistämiseen eli disambiguointiin kuitenkin kehitetään koko ajan uusia ja parempia menetelmiä[1]. FULLTEXT-projektissahan ei hyödynnetty minkäänlaista disambiguointimenetelmää, joten tässä suhteessa tuloksia voidaan varmasti parantaa.

Mielenkiintoinen kysymys on, missä vaiheessa hakemistojen kasvu tasaantuu. Kuvassa 11 (sivulla 153) näkyy, että taivutusmuotohakemiston merkkijonojen määrää kuvaava käyrä nousee alussa jyrkemmin kuin perusmuotohakemistoilla, mutta näyttää tasaantuvan, kun merkkijonojen määrä kasvaa. Teoriassa jossain vaiheessa tulee tilanne, että enimmät (käytännössä esiintyvät) taivutusmuodot ovat jo hakemistossa, joten uusia merkkijonoja syntyy lähinnä uusista erisnimistä. Tällaiset nimet jäävät perusmuoto-ohjelmalta tunnistamatta, joten perusmuotohakemistossakin nämä sananmuodot sijoittuvat tunnistamattomia sananmuotoja sisältävään taivutusmuotohakemistoon; tässä suhteessa hakemistot siis kasvavat samalla tavalla. Toisaalta osoitteiden määrä kasvaa perusmuotohakemistossa enemmän kuin taivutus-

[1] Esimerkiksi englannin kielelle EngCG-2-parseri (http://www.conexor.fi).

muotohakemistossa (kuva 12, sivulla 155). Lopputuloksena siis voisi olla, että kun tallennettavan tekstin määrä kasvaa riittävän suureksi, perusmuoto- ja taivutusmuotohakemistojen muistitilankäyttö alkavat lähetä toisiaan. Kuvassa 13 (sivulla 157) muistitilan tarvetta kuvaavat käyrät ovat kuitenkin kaikki kasvusuunnassa, joten FULLTEXT-projektin tutkimusaineiston perusteella tällaista kehitystä ei voi ennustaa, vaan tarvitaan vertailuja suuremmalla tekstimassalla.

Hypoteesi 2

> Kun ilmaisutason hakusanat esiintymätasolla korvataan vartalo-ohjelman hakusanoista tuottamilla taivutusvartaloilla, näillä automaattisesti katkaistuilla hakusanoilla saadun tulosjoukon tarkkuus on keskimäärin parempi ja saanti keskimäärin huonompi kuin tulosjoukon, joka on saatu hakijan katkaisemilla hakusanoilla. (Siis: Saanti Tn < Saanti T1 ja Tarkkuus Tn > Tarkkuus T1, missä n = 2 tai 3)

Hypoteesissa 2 väitetään, että kun hakusanat katkaistaan automaattisesti, saanti laskee. Väitteen perustana on näkemys, että merkitykseltään läheiset sanat ovat usein myös merkkijonoina lähellä toisiaan (vrt. Paice 1990). Synonyymien osaltahan tämä ei päde, eli ne ovat merkkijonoina erilaisia, mutta johdosten ja hakusanalla alkavien yhdyssanojen osalta kyllä. Automaattisen katkaisun oletetaan katkaisevan sanat niin pitkiksi, että ne eivät enää täsmää hakusanan johdosperheen muihin jäseniin.

Taulukossa 34 näkyy, että kun hakusanat katkaistaan automaattisesti (T2) tai automaattisen katkaisun lisäksi seulotaan (T3), tällaisen yhdyssanakyselyn tuottaman tulosjoukon (AC) **saanti** on sekä perus- että johdososajoukossa selvästi alempi kuin vastaavan perinteisen yhdistelmäkyselyn (T1/ABC), joka on saatu hakijan itse katkaisemilla hakusanoilla. Erot ovat tilastollisesti merkitseviä merkitsevyystasolla 0.01. Yhdyssanaosajoukossa suuntaus on sama, mutta vain seulottaessa ja yhdistettäessä hakusanat JA-operaattorilla on vertailtujen kyselytyyppien välinen ero myös tilastollisesti (ja käytännössä) merkitsevä.

Edellä olevan perusteella voidaan todeta, että hypoteesin 2 ennustama saannin lasku pätee johdoksilla. Automaattisesti katkaistut hakusanat olivat pidempiä kuin hakijan itse katkaisemat hakusanat ja jättivät pois dokumentit, jotka sisälsivät vain hakusanan johdosperheen jäseniä eikä itse hakusanaa. Perinteisessä hakutavassahan oletettiin, että hakija katkaisee hakusanat niin lyhyiksi, että johdoksetkin täsmäävät näin saatuihin merkkijonokaavioihin.

Taulukko 34. T2- ja T3-ympäristöjen yhdyssanakyselyn (AC) saanti- ja tarkkuusarvojen vertailu T1-ympäristön yhdistelmäkyselyn (ABC) saanti- ja tarkkuusarvojen kanssa perus- ja johdososajoukossa, sekä vastaavien osien yhdyssanakyselyn (ACac) ja osien yhdistelmäkyselyn (ABCabc) vertailu yhdyssanaosajoukossa. Hypoteesiin 2 sopivat joukot merkitty lihavoinnilla. Lähteinä taulukot 17 - 22.

TY	JA/ Virke	SUHTEELLINEN SAANTI				TARKKUUS			
		Joukko	**Ero**	**α**	**SJ**	**Joukko**	**Ero**	**α**	**SJ**
T2	JA	**Perus**	-8,1	**0.01**	H	**Perus**	+3,1	**0.025**	-
		Johdos	-25,6	**0.01**	O	**Johdos**	+7,8	**0.025**	H
		Yhdys	-5,0	0.1	-	**Yhdys**	+5,0	**0.05**	H
	Virke	**Perus**	-5,9	**0.01**	H	Perus	+0,7	-	-
		Johdos	-18,9	**0.01**	O	Johdos	+2,2	-	-
		Yhdys	-1,8	**-**	-	Yhdys	-3,0	-	-
T3	JA	**Perus**	-10,3	**0.01**	O	**Perus**	+4,4	**0.05**	-
		Johdos	-29,3	**0.01**	O	**Johdos**	+10,8	**0.05**	O
		Yhdys	-11,3	**0.01**	O	**Yhdys**	+11,7	**0.01**	O
	Virke	**Perus**	-7,8	**0.01**	H	Perus	+3,7	-	-
		Johdos	-22,5	**0.01**	O	Johdos	+10,9	-	-
		Yhdys	-3,5	**-**	-	Yhdys	-3,8	-	-

T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat
T3	Taivutusmuotohakemisto, seulotut hakusanat
Perus	Perusjoukko, N = 26 (T3-ympäristössä 25)
Johdos	Johdososajoukko, N = 8
Yhdys	Yhdyssanaosajoukko, N = 9
Ero	Kyselytyypin saanti- tai tarkkuusarvon ero verrattuna vastaavaan (osien) yhdistelmäkyselyn arvoon T1-ympäristössä
α	Merkitsevyystaso
SJ	Sparck Jonesin käytännön merkitsevyys: 5 – 10 prosenttiyksikön ero on huomattava (H), yli 10 prosenttiyksikön ero olennainen (O)

Tarkkuusarvojen osalta taulukossa 34 näkyy, että automaattinen katkaisu ja seulonta tuottavat kaikissa osajoukoissa perinteistä hakutapaa parempia tarkkuusarvoja, kun hakusanat on yhdistetty JA-operaattorilla. Erot ovat myös tilastollisesti merkitseviä. Virkeoperaattoria käytettäessä tulosjoukkojen tarkkuus näyttää olevan jo valmiiksi niin hyvä, ettei katkaisu enää sitä olennaisesti paranna. Hypoteesi 2 toteutui siis osittain: automaattinen katkaisu voi parantaa tarkkuutta, jos se on huono; jos tarkkuus on jo hyvä, automaattisella katkaisulla ei sitä enää paranneta. Lisäksi tarkkuus ja saanti ovat selvästi käänteisiä eli tässäkin tapauksessa tarkkuus paranee saannin

kustannuksella: lisäksi saantiarvot prosenttiyksikköinä mitaten alenevat enemmän kuin tarkkuusarvot vastaavasti prosenttiyksikköinä mitaten nousevat.

Ainoastaan yhdyssanaosajoukossa automaattinen katkaisun vaikutus oli suunnilleen sama eli saanti aleni suunnilleen yhtä monta prosenttiyksikköä kuin tarkkuus parani. Yhdyssanan osia haettaessa siis kannattaa käyttää automaattista katkaisua, jos hakusanat kytketään toisiinsa väljästi JA-operaattorilla, koska näin hakusanoista tulee pidempiä ja ne siten rajaavat osan epärelevanteista dokumenteista pois tulosjoukosta.

Hypoteesi 3

> Kun ilmaisutason hakusanat esiintymätasolla korvataan vartalo-ohjelman hakusanoista sekä näiden johdosperheistä tuottamilla taivutusvartaloilla, näillä automaattisesti katkaistuilla hakusanoilla saadun tulosjoukon tarkkuus on keskimäärin parempi ja saanti keskimäärin sama kuin tulosjoukon, joka on saatu hakijan katkaisemilla hakusanoilla. (Siis: Saanti Tn = Saanti T1 ja Tarkkuus Tn > Tarkkuus T1, missä n = 2 tai 3)

Hypoteesissa 3 siis jatkettiin hypoteesin 2 tilanteesta ja väitettiin, että kyselyn **saanti** paranee, kun johdokset lisätään kyselyyn. Näin myös kävi, kun hakusanat katkaistiin automaattisesti: kun kyselyä laajennettiin johdosperheellä, T1- ja T2-ympäristöjen (osien) yhdistelmäkyselyjen saantiarvot olivat lähes täsmälleen samat (taulukko 35). Sen sijaan seulottaessa näin ei käynyt, vaan saantiarvot jäivät perus- ja johdososajoukossa alle vertailukohteena olleen perinteisen yhdistelmäkyselyn. Tässä tapauksessa saannin aleneminen johtui kuitenkin enemmän T3-ympäristön ominaisuuksista eli seulontamenetelmän toteutustavasta. Näin ollen voidaan sanoa, että kun johdokset lisätään hakusanojen rinnalle, automaattisella katkaisulla tuotetut hakusanat antavat saanniltaan yhtä hyvän tuloksen kuin perinteiset hakijan katkaisemat hakusanat. Hypoteesi 3 siis piti saannin osalta paikkansa.

Toisaalta kyselyjen laajentaminen johdoksilla laski hakutulosten **tarkkuutta** verrattuna hypoteesin 2 tilanteeseen. Tarkkuusarvojen väliset erot eivät kuitenkaan olleet tilastollisesti merkitseviä sen enempää perusjoukossa, johdososajoukossa kuin yhdyssanaosajoukossakaan; ainoa poikkeus oli T3-ympäristön yhdyssanaosajoukko, kun hakusanat oli kytketty JA-operaattorilla. Hypoteesin 3 väite, että automaattisen katkaisun tai seulonnan tark-

Taulukko 35. T2- ja T3-ympäristöjen (osien) yhdistelmäkyselyn (ABC/ABCabc) saanti- ja tarkkuusarvojen vertailu T1-ympäristön (osien) yhdistelmäkyselyn saanti- ja tarkkuusarvojen kanssa perus- ja johdososajoukossa (yhdyssanaosajoukossa). Hypoteesiin 3 sopivat joukot merkitty lihavoinnilla. Lähteinä taulukot 17 - 22.

TY	JA/ Virke	SUHTEELLINEN SAANTI				TARKKUUS			
		Joukko	**Ero**	**α**	**SJ**	**Joukko**	**Ero**	**α**	**SJ**
T2	JA	**Perus**	-0,2	-	-	Perus	+0,7	-	-
		Johdos	0	-	-	Johdos	0	-	-
		Yhdys	0	-	-	Yhdys	+0,9	-	-
	Virke	**Perus**	-0,1	-	-	Perus	0	-	-
		Johdos	0	-	-	Johdos	0	-	-
		Yhdys	0	-	-	Yhdys	+0,3	-	-
T3	JA	Perus	-3,8	0.05	-	Perus	+1,4		-
		Johdos	-9	0.05	H	Johdos	+1,4		-
		Yhdys	-4,6	-	-	**Yhdys**	+8,3	**0.01**	H
	Virke	Perus	-2,7	0.05	-	Perus	+1,4	-	-
		Johdos	-6,3	0.05	H	Johdos	+3,6	-	-
		Yhdys	-0,6	-	-	Yhdys	+0,1	-	-

T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T2	Taivutusmuotohakemisto, Finstems-ohjelman katkaisemat hakusanat
T3	Taivutusmuotohakemisto, seulotut hakusanat
Perus	Perusjoukko, N = 26 (T3-ympäristössä 25)
Johdos	Johdososajoukko, N = 8
Yhdys	Yhdyssanaosajoukko, N = 9
Ero	Kyselytyypin saanti- tai tarkkuusarvon ero verrattuna vastaavaan (osien) yhdistelmäkyselyn arvoon T1-ympäristössä
α	Merkitsevyystaso
SJ	Sparck Jonesin käytännön merkitsevyys: 5 – 10 prosenttiyksikön ero on huomattava (H), yli 10 prosenttiyksikön ero olennainen (O)

kuus olisi parempi kuin perinteisen hakutavan tarkkuus, ei siis tällä perusteella saanut riittävästi tukea.

Koska tutkimuksen vertailuissa käytetty vartalo-ohjelma tuotti verbeistä hyvin lyhyitä vartaloita - käytännössä ei juuri sen pidempiä kuin hakijan itse katkaisemat hakusanat - tulosjoukkojen tarkkuus ei automaattisen katkaisun myötä kohentunut. Taivutusmuotohakemistossa automaattinen katkaisu ei siis tuottanut toivottua parannusta tarkkuusarvoihin.

Tulosten perusteella seulontakaan ei erityisen hyvin onnistunut parantamaan hakujen tarkkuutta. Periaatteessa seulonnan saannin pitäisi olla yhtä

hyvä kuin automaattisesti katkaistaessa (eli T2-ympäristössä) ja tarkkuuden tätä parempi. Käytännössä seulonta oli automaattista katkaisua tarkempi vain yhdyssanaosajoukossa, kun hakusanat oli yhdistetty JA-operaattorilla. Kun katkaistut yhdyssanojen osat on yhdistetty JA-operaattorilla, hakutuloksen tarkkuus siis on huono, jolloin seulonnan ja muiden tarkkuutta parantavien keinojen hyöty on ilmeinen. Muutoin seulonnan hyöty suhteessa kulutettuihin resursseihin jäi vaatimattomaksi. Perinteisillä tiedonhakutekniikoilla, kuten läheisyysoperaattoria käyttämällä, tarkkuutta ilmeisesti parannetaan helpommin kuin massiivisella hakemistosanojen seulonnalla.

Hypoteesi 4

> Kun dokumenttien sisältämät sananmuodot perusmuotoistetaan sekä tallennetaan perusmuotoisina hakemistoon ja myös hakija syöttää hakusanat perusmuodossa, näillä perusmuodoilla perusmuotohakemistosta saadun tulosjoukon tarkkuus on keskimäärin parempi mutta saanti keskimäärin huonompi kuin tulosjoukon, joka on saatu taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla. (Siis: Saanti Tn < Saanti T1 ja Tarkkuus Tn > Tarkkuus T1, missä n = 4 tai 5.)

Tutkimusympäristössä T4 ja T5 perusmuotoisia haku- ja hakemistosanoja voitiin täsmäyttää suoraan toisiinsa. Kuvaus tässä luvussa pätee myös T5-ympäristöön, vaikka tekstissä mainitaan vain T4-ympäristö; kun käytetään osittamattomia, perusmuotoisia hakusanoja, tällaiset kyselyt toimivat näissä kahdessa ympäristössä samalla tavalla.

(Osien) peruskyselyn **saantiarvot** jäivät selvästi perinteisen (osien) yhdistelmäkyselyn saantiarvoja alemmiksi kaikissa osajoukossa. Erot olivat tilastollisesti merkitseviä yleensä merkitsevyystasolla 0.01, paitsi yhdyssanaosajoukossa merkitsevyystasolla 0.05, kun hakusanat oli kytketty virkeoperaattorilla. Saannin osalta hypoteesi 4 siis pitää täysin paikkansa: jos kyselyssä käytetään vain hakusanoja perusmuodossaan, hukataan relevantteja dokumentteja. (Taulukko 36.)

T4-ympäristön peruskyselyn (tai osien peruskyselyn) **tarkkuusarvo** oli kaikissa osajoukoissa perinteisen T1-ympäristön yhdistelmäkyselyn (tai vastaavasti osien yhdistelmäkyselyn) tarkkuusarvoa korkeampi. Kun hakusanat oli kytketty toisiinsa JA-operaattorilla, tarkkuusarvojen välinen ero oli kaikissa osajoukoissa tilastollisesti merkitsevä merkitsevyystasolla 0.01. Virkeoperaattoria käytettäessä tarkkuusarvojen välinen ero oli vain johdos-

Taulukko 36. T4- ja T5-ympäristöjen (osien) peruskyselyn (A/Aa) saanti- ja tarkkuusarvojen vertailu T1-ympäristön (osien) yhdistelmäkyselyn (ABC/ ABCabc) saanti- ja tarkkuusarvojen kanssa perus- ja johdososajoukossa (yhdyssanaosajoukossa). Hypoteesiin 4 sopivat joukot merkitty lihavoinnilla. Lähteinä taulukot 23 - 28.

TY	JA/ Virke	SUHTEELLINEN SAANTI				TARKKUUS			
		Joukko	**Ero**	**α**	**SJ**	**Joukko**	**Ero**	**α**	**SJ**
T4	JA	**Perus**	-17,3	**0.01**	O	**Perus**	+7,8	**0.01**	H
		Johdos	-30,8	**0.01**	O	**Johdos**	+20,4	**0.01**	O
		Yhdys	-20,7	**0.01**	O	**Yhdys**	+23,6	**0.01**	O
	Virke	**Perus**	-13,8	**0.01**	O	Perus	+7,6	-	
		Johdos	-21,9	**0.01**	O	**Johdos**	+24,3	**0.05**	O
		Yhdys	-6,7	**0.05**	H	Yhdys	+8,8	-	
T5	JA	**Perus**	-17,3	**0.01**	O	**Perus**	+7,8	**0.01**	H
		Johdos	-30,8	**0.01**	O	**Johdos**	+20,4	**0.01**	O
		Yhdys	-20,7	**0.01**	O	**Yhdys**	+23,6	**0.01**	O
	Virke	**Perus**	-13,8	**0.01**	O	Perus	+7,6	-	
		Johdos	-21,9	**0.01**	O	**Johdos**	+24,3	**0.05**	O
		Yhdys	-6,7	**0.05**	H	Yhdys	+8,8	-	

T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat
Perus	Perusjoukko, N = 26
Johdos	Johdososajoukko, N = 8
Yhdys	Yhdyssanaosajoukko, N = 9
Ero	Kyselytyypin saanti- tai tarkkuusarvon ero verrattuna vastaavaan (osien) yhdistelmäkyselyn arvoon T1-ympäristössä
α	Merkitsevyystaso
SJ	Sparck Jonesin käytännön merkitsevyys: 5 – 10 prosenttiyksikön ero on huomattava (H), yli 10 prosenttiyksikön ero olennainen (O)

osajoukossa tilastollisesti merkitsevä (merkitsevyystasolla 0.05); perus- ja yhdyssanaosajoukossa ero ei ollut tilastollisesti merkitsevä.

Hypoteesin 4 väite, että perusmuotoisilla hakusanoilla saadaan paremmat tarkkuusarvot kuin perinteisesti itse katkaistuilla hakusanoilla, siis sai tukea erityisesti, kun hakusanat oli kytketty JA-operaattorilla. Virkeoperaattorilla hypoteesi sai tukea varsinaisesti vain johdososajoukossa - virkeoperaattorin käyttö siis parantaa tarkkuutta jo sen verran, että perusmuotojen käyttö ei enää olennaisesti lisää tarkkuutta.

Jos siis taivutusmuotohakemiston ja katkaistujen hakusanojen käyttöön tottunut hakija hakee perusmuotohakemistosta ja haluaa tulosjoukon saannin olevan hyvän, ei riitä, että hän hakee vain hakusanojen perusmuodoilla. Tällöin kyselyn ala jää selvästi suppeammaksi kuin vastaavan perinteisen kyselytavan. Tämän tutkimuksen vertailujen perusteella saanti laskee useampia prosenttiyksikköjä kuin mitä tarkkuus samalla prosenttiyksikköinä mitaten kasvaa.

Hypoteesi 5

> Kun dokumenttien sananmuodot palautetaan perusmuotoon sekä tallennetaan perusmuotoisina hakemistoon, ja myös hakijan syöttämät hakusanat johdosperheineen ovat perusmuodossa, tällä johdosperheellä perusmuotohakemistosta saadun tulosjoukon tarkkuus on keskimäärin parempi, mutta saanti keskimäärin sama kuin tulosjoukon, joka on saatu taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla. (Siis: Saanti Tn = Saanti T1 ja Tarkkuus Tn > Tarkkuus T1, missä n = 4 tai 5.)

Hypoteesissa 5 edellisen hypoteesin asetelmaa muutetaan siten, että kyselyyn lisätään hakusanojen rinnalle myös niiden johdosperheen jäsenet. Tämä voidaan toteuttaa esimerkiksi poimimalla johdosperhe (puoli)automaattisesti hakutesauruksesta. Hypoteesin 5 mukaan tämä nostaa saantiarvot samalle tasolle kuin hakijan katkaisemia hakusanoja käytettäessä.

Tulosten perusteella näin ei kuitenkaan käynyt (taulukko 37). Johdosperheen lisääminen paransi **saantia** verrattuna siihen, että olisi käytetty vain hakusanojen perusmuotoja. Saantiarvot jäivät silti edelleen alle perinteisen yhdistelmäkyselyn saantiarvojen. Erot perinteiseen hakutapaan olivat myös tilastollisesti merkitsevät - tästä poikkeuksena vain yhdyssanaosajoukko, kun hakusanat oli kytketty toisiinsa virkeoperaattorilla. Tämä johtunee siitä, että virkeoperaattoria käytettäessä saantiarvot jäävät muutenkin alhaisiksi, jolloin eri vaihtoehtojen väliset erot eivät muutenkaan ole suuret. Näin ollen hypoteesi 5 ei saannin osalta saa tukea: pelkkä johdosperheen lisääminen kyselyyn ei paranna saantia yhdistelmäkyselyn tasolle.

Tarkkuusarvot olivat selvästi paremmat kuin perinteisen yhdistelmäkyselyn, joten tästä osin hypoteesin 5 voidaan sanoa pitävän paikkaansa, vaikka tarkkuusarvot olivatkin alemmat kuin peruskyselyn tarkkuusarvot.

Taulukko 37. T4- ja T5-ympäristöjen (osien) johdoskyselyn (AB/ABab) saanti- ja tarkkuusarvojen vertailu T1-ympäristön (osien) yhdistelmäkyselyn (ABC/ ABCabc) saanti- ja tarkkuusarvojen kanssa perus- ja johdososajoukossa (yhdyssanaosajoukossa). Hypoteesiin 5 sopivat joukot merkitty lihavoinnilla. Lähteinä taulukot 23 - 28.

TY	JA/ Virke	SUHTEELLINEN SAANTI				TARKKUUS			
		Joukko	**Ero**	**α**	**SJ**	**Joukko**	**Ero**	**α**	**SJ**
T4	JA	Perus	-10,3	0.01	O	**Perus**	+4,4	**0.05**	-
		Johdos	-8,2	0.01	H	**Johdos**	+9,6	**0.05**	H
		Yhdys	-15,1	0.01	O	**Yhdys**	+14,0	**0.01**	O
	Virke	Perus	-8,1	0.01	H	Perus	+3,8	-	
		Johdos	-3,4	0.05	-	**Johdos**	+12,1	**0.05**	O
		Yhdys	-4,3	-		Yhdys	+9,1	-	
T5	JA	Perus	-10,3	0.01	O	**Perus**	+4,4	**0.01**	-
		Johdos	-8,2	0.01	H	**Johdos**	+9,6	**0.05**	H
		Yhdys	-15,1	0.01	O	**Yhdys**	+14,0	**0.01**	O
	Virke	Perus	-8,1	0.01	H	Perus	+3,8	-	
		Johdos	-3,4	0.05	-	**Johdos**	+12,1	**0.05**	O
		Yhdys	-4,3	-		Yhdys	+9,1	-	

T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat
Perus	Perusjoukko, N = 26
Johdos	Johdososajoukko, N = 8
Yhdys	Yhdyssanaosajoukko, N = 9
Ero	Kyselytyypin saanti- tai tarkkuusarvon ero verrattuna vastaavaan (osien) yhdistelmäkyselyn arvoon T1-ympäristössä
α	Merkitsevyystaso
SJ	Sparck Jonesin käytännön merkitsevyys: 5 – 10 prosenttiyksikön ero on huomattava (H), yli 10 prosenttiyksikön ero olennainen (O)

T5-ympäristössä saatiin yleisesti ottaen samanlaiset tulokset kuin T4-ympäristössä. Perusjoukossa tarkkuusarvojen merkitsevyydessä oli yksi ero, kun kyselyssä oli käytetty JA-operaattoria: T5-ympäristössä arvojen välinen ero oli merkitsevä merkitsevyystasolla 0.01, T4-ympäristössä vasta alemmalla merkitsevyystasolla 0.05.

On kuitenkin huomattava, että vaikka johdosten lisääminen kyselyyn ei täysin johtanutkaan hypoteesin 5 esittämään asetelmaan, tilanne on silti parempi kuin pelkästään hakusanan perusmuotoja käytettäessä. Johdoskyselyn saanti on selvästi peruskyselyn saantia parempi (ero on tilastollisesti merkit-

sevä), mutta tarkkuus vain hieman huonompi (eikä tuo vähäinen ero ole myöskään tilastollisesti merkitsevä). Tällä perusteella T4- ja T5-ympäristöissä ei kannata hakea vain perusmuodoilla, vaan laajentaminen johdosperheellä on suositeltavaa.

Kyselyn laajentaminen yhdyssanoilla

Kyselyn laajentaminen johdosperheellä ei siis täysin vastannut hypoteesin 5 väitteitä. Toinen vaihtoehto on laajentaa peruskyselyä johdosperheen sijasta yhdyssanoilla. T4-ympäristössä yhdyssanat voidaan automaattisesti lisätä kyselyyn siten, että hakusana syötetään ensin perusmuodossa vartaloohjelmalle, jonka tuottamat vartalot lisätään kyselyyn hakusanan rinnalle. T5-ympäristössä puolestaan hakusanaan lisätään automaattisesti yhdyssanan katkaisukohtia osoittavat katkaisumerkit, jonka jälkeen nämä yhdyssanan osat lisätään kyselyyn hakusanan rinnalle.

T4- ja T5-ympäristöissä sekä perusjoukossa että johdososajoukossa yhdyssanakyselyjen **saantiarvot** jäivät alemmiksi kuin perinteisen yhdistelmäkyselyn. Saantiarvojen väliset erot olivat sekä JA- että virkeoperaattoria käytettäessä tilastollisesti merkitseviä merkitsevyystasolla 0.01. Toisaalta, kun vertailukohtana on lähtötilanne eli peruskysely, yhdyssanakyselyn saanti oli perusjoukossa huomattavasti parempi kuin peruskyselyn, ja johdososajoukossakin jonkin verran parempi. (Taulukko 38.)

Peruskyselyn laajennus yhdyssanakyselyksi kuitenkin alentaa **tarkkuutta** - perusjoukossa useampia prosenttiyksikköjä vähemmän ja johdososajoukossa useampia prosenttiyksikköjä enemmän kuin saanti prosenttiyksikköinä mitaten paranee. Tarkkuusarvot olivat kuitenkin korkeammat kuin perinteisen yhdistelmäkyselyn. JA-operaattoria käytettäessä erot olivat myös tilastollisesti merkitsevät. Näin ollen perusjoukossa ja johdososajoukossa sopivan kyselytyypin valinta riippuu siitä, haluaako hakija painottaa enemmän saantia vai tarkkuutta.

Sen sijaan yhdyssanaosajoukossa osien yhdyssanakyselyn ja perinteisen osien yhdistelmäkyselyn saantiarvojen väliset erot eivät olleet tilastollisesti merkitseviä. Tässä tapauksessa siis hakusanalla alkavien yhdyssanojen lisääminen kyselyyn nostaa saannin lähes yhdistelmäkyselyn tasalle. Tämä tulos saatiin sekä T4- että T5-ympäristössä.

Taulukko 38. T4- ja T5-ympäristöjen (osien) yhdyssanakyselyn (AC/ACac) saanti- ja tarkkuusarvojen vertailu T1-ympäristön (osien) yhdistelmäkyselyn (ABC/ABCabc) saanti- ja tarkkuusarvojen kanssa perus- ja johdososajoukossa (yhdyssanaosajoukossa). Hypoteesin 5 periaatteeseen sopivat joukot merkitty lihavoinnilla. Lähteinä taulukot 23 - 28.

TY	JA/ Virke	SUHTEELLINEN SAANTI				TARKKUUS			
		Joukko	**Ero**	**α**	**SJ**	**Joukko**	**Ero**	**α**	**SJ**
T4	JA	Perus	-7,9	0.01	H	**Perus**	+5,0	**0.01**	H
		Johdos	-25,0	0.01	O	**Johdos**	+10,3	**0.01**	O
		Yhdys	-4,4	-		**Yhdys**	+7,4	**0.05**	H
	Virke	Perus	-5,7	0.01	H	Perus	+2,6	-	
		Johdos	-18,2	0.01	O	Johdos	+8,4	**-**	
		Yhdys	-1,2	-		Yhdys	+2,6	-	
T5	JA	Perus	-6,3	0.05	H	**Perus**	+5,2	**0.01**	H
		Johdos	-24,5	0.01	O	**Johdos**	+8,0	**0.01**	H
		Yhdys	-3,7	-		Yhdys	+9,0	0.1	H
	Virke	Perus	-5,5	0.01	H	Perus	+2,6	-	
		Johdos	-18,2	0.01	O	Johdos	+7,5	**-**	
		Yhdys	+1,3	-		Yhdys	+1,4	-	

T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat
Perus	Perusjoukko, N = 26
Johdos	Johdososajoukko, N = 8
Yhdys	Yhdyssanaosajoukko, N = 9
Ero	Kyselytyypin saanti- tai tarkkuusarvon ero verrattuna vastaavaan (osien) yhdistelmäkyselyn arvoon T1-ympäristössä
α	Merkitsevyystaso
SJ	Sparck Jonesin käytännön merkitsevyys: 5 – 10 prosenttiyksikön ero on huomattava (H), yli 10 prosenttiyksikön ero olennainen (O)

Osien yhdyssanakyselyn tarkkuusarvo taas oli kummankin tutkimusympäristön yhdyssanaosajoukossa perinteisen T1-ympäristön osien yhdistelmäkyselyn tarkkuusarvoa korkeampi. Kun hakusanat oli kytketty toisiinsa JA-operaattorilla, tarkkuusarvojen välinen ero oli T4-ympäristössä tilastollisesti merkitsevä, T5-ympäristössäkin lähes merkitsevä. Virkeoperaattoria käytettäessä erot eivät olleet tilastollisesti merkitseviä.

Yhdyssanaosajoukossa siis kannattaa laajentaa kyselyä hakusanan sisältäviin yhdyssanoihin, jolloin tilanne lähes vastaa hypoteesin 5 tilannetta, eli

saanti on lähes sama kuin perinteisen yhdistelmäkyselyn, ja tarkkuus on tätä parempi.

Hypoteesissa 3 asetelma oli muuten sama kuin hypoteesissa 5, mutta siinä kyselyt tehtiin taivutusmuotohakemistosta. Automaattisen katkaisun tuottamien tulosjoukkojen tarkkuus ei taivutusmuotohakemistossa ollut parempi kuin itse katkaistuja hakusanoja käytettäessä. Sen sijaan perusmuotohakemistossa automaattinen katkaisu toimi toivotulla tavalla: vartalo-ohjelman tuottamilla vartaloilla saadut hakutulokset olivat tarkempia kuin hakijan itse katkaisemilla hakusanoilla saadut, eikä saanti laskenut yhtä monta prosenttiyksikköä kuin tarkkuus parantui. Automaattinen katkaisu siis toimii perusmuotohakemistossa tarkemmin kuin taivutusmuotohakemistossa.

Hypoteesi 6

> Kun hakija tekee perusmuotohakemistosta kyselyn, jossa hakusanat ja hakusanojen johdosperheen jäsenet ovat perusmuodossa ja lisäksi hakusana ja tämän johdosperheen jäsenet katkaistaan automaattisesti näillä sanoilla alkavien yhdyssanojen löytämiseksi, tällaisen kyselyn tuloksena saadun tulosjoukon tarkkuus on keskimäärin parempi ja saanti keskimäärin parempi kuin tulosjoukon, joka on saatu taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla. (Siis: Saanti T4 > Saanti T1 ja Tarkkuus T4 > Tarkkuus T1.)

Hypoteesi 7

> Kun hakija tekee ositetusta perusmuotohakemistosta kyselyn, jossa hakusanat, hakusanojen johdosperheen jäsenet ja yhdyssanojen osat ovat perusmuodossa ja yhdyssanan osina, tällaisen kyselyn tuloksena saadun tulosjoukon tarkkuus on keskimäärin huonompi, mutta saanti keskimäärin parempi kuin tulosjoukon, joka on saatu taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla; lisäksi tällaisen kyselyn saanti on keskimäärin parempi kuin (osittamattomasta) perusmuotohakemistosta saadun tulosjoukon saanti. (Siis: Saanti T5 > Saanti T4 > Saanti T1 ja Tarkkuus T5 < Tarkkuus T1.)

Kun T4-ympäristön (osien) peruskysely laajennettiin sekä johdosperheellä että yhdyssanoilla (osien) yhdistelmäkyselyksi, tulosjoukkojen **saantiarvo** nousi vastaavan T1-ympäristön kyselyn saantiarvon tasalle, johdos- ja yhdyssanaosajoukossa jopa hivenen sen yli. Saantiarvojen väliset erot eivät kuitenkaan olleet tilastollisesti merkitseviä. Tältä osin tulos ei riittävästi tue hypoteesia 6.

Taulukko 39. T4- ja T5-ympäristöjen (osien) yhdistelmäkyselyn (ABC/ABCabc) saanti- ja tarkkuusarvojen vertailu T1-ympäristön (osien) yhdistelmäkyselyn saanti- ja tarkkuusarvojen kanssa perus- ja johdososajoukossa (yhdyssanaosajoukossa). Lähteinä taulukot 23 - 28.

TY	JA/ Virke	SUHTEELLINEN SAANTI				TARKKUUS			
		Joukko	Ero	α	SJ	Joukko	Ero	α	SJ
T4	JA	Perus	0	-		Perus	+2,3	0.1	
		Johdos	+0,7	-		Johdos	+1,3	-	
		Yhdys	+0,5	-		Yhdys	+2,5	-	
	Virke	Perus	+0,1	-		Perus	+0,7	-	
		Johdos	+0,7	-		Johdos	+2,5	-	
		Yhdys	+0,6	-		Yhdys	+2,6	-	
T5	JA	Perus	+3,1	-		Perus	+3,0	0.1	
		Johdos	+6,1	-		Johdos	+0,9	-	
		Yhdys	+10,3	-		Yhdys	+4,5	-	
	Virke	Perus	+0,9	-		Perus	+0,7	-	
		Johdos	+2,8	-		Johdos	+1,1	-	
		Yhdys	+8,9	-		Yhdys	+1,0	-	

T1	Taivutusmuotohakemisto, hakijan katkaisemat hakusanat
T4	Perusmuotohakemisto ja perusmuodossa annetut hakusanat
T5	Ositettu perusmuotohakemisto ja perusmuodossa annetut hakusanat
Perus	Perusjoukko, N = 26
Johdos	Johdososajoukko, N = 8
Yhdys	Yhdyssanaosajoukko, N = 9
Ero	Kyselytyypin saanti- tai tarkkuusarvon ero verrattuna vastaavaan (osien) yhdistelmäkyselyn arvoon T1-ympäristössä
α	Merkitsevyystaso
SJ	Sparck Jonesin käytännön merkitsevyys: 5 – 10 prosenttiyksikön ero on huomattava (H), yli 10 prosenttiyksikön ero olennainen (O)

Kun vastaavat kyselyjen laajennukset tehtiin T5-ympäristössä, saantiarvot siinäkin nousivat kaikissa osajoukoissa korkeammalle kuin T1-ympäristön (osien) yhdistelmäkyselyllä. T5-ympäristön saantiarvot olivat myös korkeammat kuin T4-ympäristössä, joten tulokset sinänsä olivat hypoteesin 7 mukaiset - koska erot eivät kuitenkaan olleet tilastollisesti merkitseviä, riittävää tukea hypoteesille ei saatu.

(Osien) yhdistelmäkyselyjen **tarkkuusarvot** olivat T4-ympäristössä korkeammat kuin perinteisen T1-ympäristön tarkkuusarvot. Vaikka T4- ja T1-ympäristöjen tarkkuusarvojen välille ei löytynyt tilastollisesti merkitsevää eroa (vain tasolla 0.1 perusjoukossa), T4-ympäristön paremmuus verrattuna

T1-ympäristöön oli kuitenkin systemaattinen (T4 ≥ T1). Koska ero ei ollut tilastollisesti merkitsevä, hypoteesin 6 väitteelle paremmasta tarkkuudesta ei kuitenkaan saatu riittävästi tukea.

Myös T5-ympäristössä tarkkuusarvot olivat keskimäärin korkeammat kuin T1-ympäristössä, vaikka hypoteesissa 7 oletettiin käyvän toisin päin. Erot eivät kuitenkaan olleet tilastollisesti merkitseviä. Tulosten perusteella voidaan kuitenkin todeta, että yhdyssanojen keski- ja loppuosilla hakeminen ei näytä huonontavan tarkkuutta, koska ositetusta perusmuotohakemistosta hakeminen on täsmällisempää kuin haettaessa taivutusmuotohakemistosta katkaistuilla hakusanoilla.

Ositetun perusmuotohakemiston voi kuitenkin sanoa olevan taivutusmuotohakemistoa parempi siinä suhteessa, että perusmuodot ovat hakijan kannalta selkeämpiä käyttää kuin katkaistut hakusanat. Lisäksi siitä löytyvät yhdyssanojen keski- ja loppuosat helposti. Hakijalla on myös hyvät mahdollisuudet muunnella kyselyä sen mukaan, haluaako hän painottaa saantia vai tarkkuutta. Hakijan on kuitenkin muistettava, että hyvä saanti edellyttää, että johdokset ja yhdyssanat on otettu kyselyssä huomioon.

Edelläkuvattujen vertailujen jälkeen voidaan tietenkin pohtia, onko vertailujen lähtökohta oikea. Perusoletuksenahan oli, että ammattilaishakija pyrkii saamaan johdokset kyselyihin mukaan mahdollisimman kattavasti. Vaikka ammattilaishakijan ammattitaitoon kuuluukin hakusanan eri muotojen ja vaihtoehtoisten rinnakkaistermien arvioiminen ja sopivien hakuavainten lisääminen kyselyyn tarvittaessa, ei kaikissa kyselyissä pyritä maksimaaliseen saantiin. Jos perinteisen hakutavan lähtöoletuksia lievennetään ja johdoksia painotetaan vähemmän kuin tässä tutkimuksessa, voidaan perusmuotohakemiston ja ositetun perusmuotohakemiston yhdyssanakyselyllä (AC) ja osien yhdyssanakyselyllä (ACac) päästä melko lähelle taivutusmuotohakemistosta tehdyn perinteisen yhdistelmäkyselyn (ABC) tai osien yhdistelmäkyselyn (ABCabc) tuloksia.

Toisaalta johdokset ovat tärkeitä, jos kyselyn saannin halutaan olevan mahdollisimman hyvä. Perinteisesti haettaessa ne usein tulevat tulosjoukkoon mukaan ikään kuin sivutuotteena. Näin ei kuitenkaan aina tapahdu. Jos hakusanana on autoverotus ja hakija katkaisee sen merkkijonokaavioksi *autoverotu**, löytymättä jäävät sellaiset dokumentit, joissa esiintyy pelkästään autovero eri muodoissaan. Jos siis johdoksia pidetään tärkeinä ja ne

kaikki halutaan löytää, johdoksia pystytään (ositetussa) perusmuotohakemistossa hakemaan täsmällisemmin kuin perinteisessä hakemistossa. Saanti pysyy samana tai jopa paranee verrattuna perinteiseen hakutapaan, ja kyselyjen tarkkuus on parempi kuin katkaistuja hakusanoja käytettäessä.

11.3 HOMOGRAFIASTA JOHTUVAT ONGELMAT

Usein homografisten ilmausten perusmuodoista yksi on toisia huomattavasti yleisempi eli todennäköisin tulkinta. Esimerkiksi sanomalehtitekstissä on todennäköisempää, että puhelin-sana viittaa laitteeseen kuin että se olisi puhella-verbin yksi taivutusmuoto. Perusmuoto-ohjelman sanakirja todennäköisesti sisältää yleisimmän homografin perusmuodon, mutta harvinaisten homografien perusmuodot ovat mukana satunnaisesti. Sanakirjasta puuttuvat sanat joko eivät päädy perusmuotohakemistoon tai ne joutuvat sinne väärässä muodossa. Harvinaisten sanojen saanti siis voi laskea, jos sananmuodot vain perusmuotoistetaan tallennusvaiheessa eikä hakuvaiheessa yritetä korjata mahdollisia virhetulkintoja.

FULLTEXT-projektin useissa ongelmakyselyissä perusmuotohakemistosta löydettiin vähemmän dokumentteja kuin perinteisellä tavalla haettaessa. Hukatut dokumentit olivat vain yhdessä tapauksessa (*Sri Lanka*) myös relevantteja. Koska testikyselyitä ei ollut paljon ja niistäkin osa oli keksittyjä, tämän tutkimuksen tulosten perusteella ei voida arvioida, miten merkittävä ongelma todellisuudessa on. Huonontavatko homografit saantiarvoja myös käytännössä vai onko kyseessä lähinnä teoreettinen riski?

Perinteisesti homografit ovat olleet ongelma paremminkin tarkkuuden kuin saannin kannalta. Esimerkiksi *Salmen*-haussa saatiin hakijan katkaisemalla hakusanalla tulokseksi 36 dokumenttia, joista todellisuudessa vain seitsemässä mainittiin *Leif Salmén*. Useimmissa dokumenteissa esiintyi salmi-sanan genetiivi.

Lancaster (1986, s. 69) kuitenkin toteaa, että homografia ei ole niin suuri ongelma kuin ensisilmäyksellä olettaisi. Hänen mukaansa homografeista koituu ongelmia silloin, kun haetaan vain yhdellä, irrallisella hakusanalla. Käytännössä kyselyt sisältävät useampia konjunktiivisia hakusanoja, jolloin kyselyn muut hakusanat tavallisesti rajaavat pois dokumentit, joissa homografi on väärän sanan esiintymä. Tähän näkemykseen päätyi myös homo-

grafien vaikutusta tekstihaun tuloksiin tutkinut Leppänen (1996). Hänen mukaansa homografia ei käytännössä osoittautunut erityiseksi ongelmaksi.

Leppäsen (1996) tutkimuksessa kuitenkin käytettiin perinteistä taivutusmuotohakemistoa. Perusmuotohakemistossa ongelman luonne on toisenlainen: siinäkään homografit sinänsä eivät ole suuri ongelma, mutta **väärin tulkitut** homografit ovat. Taivutusmuotohakemistosta *Salmen*-kyselyllä saaduista 36 dokumentista relevantteja oli 8, homografeja (kuten salmi-sanan genetiivejä) sisältäviä 14, loput yhdyssanoja. Perusmuotohakemistossa taas Salmén-nimi tulkittiin salmi- ja Salme-sanojen esiintymäksi, koska sitä ei löytynyt perusmuoto-ohjelman sanakirjasta. Jotta saanti ei olisi jäänyt huonommaksi kuin perinteisellä tavalla haettaessa, oli salmen-sana haettava kahdesta hakemistosta: tunnistamatta jääneet muodot taivutusmuotohakemistosta ja väärintulkitut muodot perusmuotohakemistosta. Näin tehden saatiin tulokseksi kokonaista 166 dokumenttia, joista 149 kappaletta oli salmi-sanan tuottamia osumia perusmuotohakemistosta. Näistä dokumenteista kuitenkin vain osassa esiintyi Salmén-nimen taivutusmuotohomografi eli salmi-sanan yksikön genetiivi. Muissa dokumenteissa oli aivan muu, alunperin ei-homografinen salmi-sanan taivutusmuoto.

Jos siis tietty sana puuttuu perusmuoto-ohjelman sanakirjasta ja sen sananmuoto tulkitaan väärin (taivutusmuoto)homografikseen, kyselyn saanti voi kärsiä, mikäli ei tehdä korjauskyselyä eli etsitä perusmuotohakemistosta myös mahdollisia väärintulkittuja muotoja. Tällainen korjauskysely puolestaan voi huonontaa kyselyn tarkkuutta, mikäli ei tehdä toista korjauskyselyä. Luvussa 10 kuvattiin menetelmä, jossa taivutusvartaloiden avulla tehdään uusi korjauskysely ja lopputuloksena saadaan samat dokumentit kuin itse katkaistuja hakusanoja käytettäessä. Tarkkuuskin voidaan siis perusmuotohakemistossa saada virhetulkinnoista huolimatta samalle tasolle kuin taivutusmuotohakemistosta haettaessa.

Korjaushakuja kuitenkin kannattanee käyttää vain silloin, kun kyselyssä ei ole muita, rajaavia hakuavaimia, jotka rajaavat väärät osumat pois. Tarkkuutta korjaava kysely tehtäisiin vain silloin, kun tulosjoukon koko muuten jäisi liian suureksi. FULLTEXT-projektissakin todettiin, että konjunktiiviset hakusanat rajaavat homografien virhetulkinnat tehokkaasti pois: Kun T4-ympäristössä haettiin Inga Sulin -nimeä, pelkällä *sulin*-hakusanalla saatiin yhteensä 171 osumaa perusmuoto- ja tunnistamattomien sananmuotojen hakemistoista (Sulin -> sulka, sulaa, sula). Kun tämän hakusanan rinnalle

kyselyyn sitten lisättiin rajaava *inga*-hakusana, se karsi ylimääräiset osumat pois ja kyselyn lopputuloksena saatiin samat kaksi relevanttia dokumenttia kuin perinteisellä tavalla hakijan itse katkaisemilla hakusanoilla haettaessa.

Toinen esimerkki homografian aiheuttamasta huonosta tarkkuudesta oli halva (hakupyyntö 43). Halva-perusmuoto sisältyi Morfo-ohjelman sanakirjaan. Sen monet taivutusmuodot, kuten genetiivi halvan, kuitenkin ovat identtisiä paljon yleisemmin esiintyvän halpa-adjektiivin taivutusmuotojen kanssa. Hakijan itse katkaisema hakusana *halva** tuotti taivutusmuotohakemistosta 130 dokumenttia. Kun hakusana katkaistiin automaattisesti (T2), Hahmotin lisäsi kyselyyn *halva*-vartalon rinnalle myös *halvo*-vartalon. Näillä kahdella vartalolla haettaessa saatiin seitsemän dokumenttia enemmän kuin hakijan katkaisemilla hakusanoilla haettaessa eli yhteensä 137 kappaletta. Finstems-ohjelma puolestaan tuotti *halvoi-* ja *halvoj*-vartalot, joilla haettaessa saatiin viisi dokumenttia enemmän kuin itse katkaistuilla hakusanoilla, siis yhteensä 135 dokumenttia. Perusmuotohakemistosta haettaessa perusmuodossa syötetty *halva*-hakusana tuotti 95 osumaa.

Kun *halva*-kyselyn tuloksena saatuja dokumentteja tarkasteltiin tarkemmin, ei niistä yhdenkään aiheena ollut *halva*. Kyselyn tarkkuus siis oli huono (olematon), suoritettiinpa kysely sitten taivutus- tai perusmuotohakemistosta. Dokumenteista 95 (eli ne, jotka löytyivät perusmuotohakemistosta) sisälsivät sananmuodon, joka oli homografinen jonkin halva-sanan taivutusmuodon kanssa. Tavallisesti kyseessä oli jokin halpa-sanan taivutusmuoto. Perusmuotohakemistosta kuitenkin saatiin vähemmän dokumentteja kuin taivutusmuotohakemistosta haettaessa. Finstems- tai Hahmotin-ohjelmalla katkaistut hakusanat palauttivat samat 95 dokumenttia kuin edellä ja niiden lisäksi vielä dokumentteja, joissa esiintyi hakusanan kanssa samalla tavalla alkava sana (halvaus, halvaantua, Halvari ym.).

Kun teksteihin sisältyviä sananmuotoja käsitellään vain sanatasolla, monet ilmaukset pakostikin jäävät monitulkintaisiksi ja hakemistoihin joutuu paljon homografisia sananmuotoja. Ratkaisu tai ainakin helpotus ongelmaan ovat sellaiset luonnollisen kielen tulkintaohjelmat, jotka pystyvät ottamaan myös sanojen tekstiyhteyden huomioon. Monitulkintaisuuksia karsivan disambiguointiohjelman avulla voidaan ratkaista esimerkiksi, onko voi substantiivin vai verbin esiintymä. FULLTEXT-tutkimuksen aikaan tällaista ohjelmaa ei ollut saatavilla suomen kieltä varten, sen sijaan englannin kielelle sellainen oli jo kehitetty (Karetnyk et al. 1991; Voutilainen 1994; Conexor

2000). Kun automaattista disambiguointimenetelmää ei ole käytettävissä, on monitulkintaisia sananmuotoja sisältävät dokumentit tulostettava hakijan luettavaksi, joka sitten itse päättelee homografin oikean tulkinnan tekstiyhteyden perusteella.

11.4 SULKUSANALISTA

Mikäli hakujärjestelmässä on tallennusvaiheessa käytetty sulkusanalistaa, se pitäisi myös hakuvaiheessa käydä vastaavassa vaiheessa läpi. Tämä periaate pätee sekä taivutusmuoto- että perusmuotohakemistoon. Tämän tutkimuksen testihakuja ei tarkistettu sulkusanalistasta, koska tiedettiin, ettei mikään testikyselyissä käytetyistä hakusanoista ollut tällä listalla.

Usein kaupalliset suorakäyttöjärjestelmät dokumentoivat asiakkailleen, millaisia sulkusanalistoja tallennusvaiheessa on käytetty. Hakuvaiheessa sulkusanalistojen vaikutus hakuun ei yleensä kuitenkaan käy ilmi. Jos hakija käyttää sanaa, joka on sulkusanalistalla, tuloksena on tyhjä tulosjoukko. Tällaisessa tapauksessa hakujärjestelmät harvemmin ilmoittavat syytä siihen, miksi mitään ei löydy, vaan hakijan on huomattava itse, että hakusana on sulkusanalistalla ja siksi puuttuu hakemistosta.

Taivutusmuotohakemiston sulkusanalistan pitäisi sisältää pois suljettavan sanan kaikki mahdolliset taivutusmuodot (ellei sitten sulkusanalistalla ole vain taipumattomia partikkeleita). Täydellisen sulkulistan laatiminen suomen kielen taipuville sanoille on käytännössä mahdotonta, substantiiveillahan on taivutusmuotoja teoriassa noin 2 000 kappaletta ja muissa sanaluokissa vieläkin enemmän (Koskenniemi 1985a). Jos esimerkiksi olla-verbin taivutusmuoto olin on sulkusanalistalla, pitäisi myös muotojen olinkin, olinhan jne. olla listalla mukana tai ne tallentuvat hakemistoon.

Sulkusanalistan laatijan pitäisi myös varoa, ettei hän tiettyä sananmuotoa karsiessaan samalla sulje pois sen kanssa homografista toisen sanan sananmuotoa. Esimerkkitapauksessa myös Olin-nimi karsiutuu olin-verbin lisäksi eikä siten kysely *Tenho Olinista* hakusanalla *olin* löydä tuota sanaa hakemistosta, vaikka se olisikin itse dokumenteissa esiintynyt. Toisaalta katkaistu hakusana *olin** voi tuottaa sulkusanalistan ohittaneita sananmuotoja, mikäli sulkusanalistan laatija ei ole ottanut listalle kaikkia mahdollisia taivutusmuotoja (esimerkiksi unohtanut listalta taivutusmuodot Olinin tai Olinkin).

Myöskään perusmuotohakemistoissa sulkusanalistojen käyttäminen ei ole ongelmatonta. BASIS-K-versiossa hakemistoon tulevat sananmuodot oli mahdollista perusmuotoistaa vasta sen jälkeen, kun sulkusanalista oli käyty läpi. Tällöin voi käydä niin, että ei-toivottu sana joutuu hakemistoon, vaikka se olisikin sulkusanalistalla. Oletetaan esimerkiksi, että olla-sanan eri muodot ovat sulkusanalistalla, ja ne karsitaan alkuvaiheessa pois merkkijonojen vertailun perusteella. Toisaalta olin, olinkin yms. on jätetty sulkusanalistalta pois, koska ne halutaan hakemistoon mukaan mahdollisina Olin-nimen esiintyminä. Kun nämä muodot ovat päässeet sulkusanalistasta läpi, perusmuoto-ohjelma tuottaa niistä sekä olin-substantiivin että olla-verbin (olettaen että nämä molemmat sanat sisältyvät sen sanakirjaan). Olla-verbi on huomattava poistaa myös tässä vaiheessa, tai muuten se joutuu hakemistoon.

Jos BASIS-K-järjestelmän käyttäjä hakuvaiheessa antaisi hakusanaksi olla, hakujärjestelmän pitäisi heti havaita, että olla esiintyy sulkusanalistalla eikä jatkaa hakua pidemmälle. Muuten hakemistosta löydetään vain ne olla-hakemistosanat, jotka esiintyivät alkuperäistekstissä jossain muussa kuin sulkusanalistalla määritellyssä taivutusmuodossa. Tulokseksi saataisiin siis vain hyvin pieni joukko olla-verbin esiintymistä. Koska kyselyn tuloksena ei ollut tyhjä joukko, hakija ei välttämättä aavista, että jotain on vialla. Automaattista virheenkorjaustakaan ei tehdä, jos se käynnistyy vain silloin, kun kyselyn tuloksena on tyhjä joukkko.

Edellä käytetty esimerkkisana ei ole paras mahdollinen, koska käytännössä tuskin kukaan käyttää hakusanana olla-verbiä. Se kuitenkin havainnollistaa sen, miksi sulkusanalistaa pitäisi soveltaa suomenkieliseen tekstiin varoen ja erityisesti silloin, kun käsiteltävät sanat sulkusanalistan soveltamisen jälkeen palautetaan perusmuotoon. Mikäli sulkusanalista sisältää vain taipumattomia sanoja tai mikäli sulkusanalistaa sovelletaan vasta perusmuotoistamisen jälkeen, edellä kuvatut ongelmat vältetään. Joka tapauksessa sulkusanalistan käyttö tulisi suunnitella huolellisesti, jotta hakijaa ei pakoteta ylen mutkikkaaseen ajatusakrobatiaan.

Mikäli sulkusanalistaa käytetään, hakusanan tarkistaminen pitäisi aloittaa tutkimalla, onko hakusana sulkusanalistalla. Mikäli hakusana löytyy listalta, tämä pitäisi ilmoittaa käyttäjälle eikä jatkaa kyselyn suorittamista. Näin hakija saa mahdollisuuden keksiä vaihtoehtoisia hakusanoja tai kohdistaa kyselyn suoraan dokumenttien tekstiin, joissa sanat esiintyvät alkuperäises-

sä muodossaan. Mikäli hakusana ei ole sulkusanalistalla, syötteen muut tarkistukset toteutetaan sen mukaisesti, miten suomen kielen tulkintaohjelmia on tallennusvaiheessa sovellettu.

11.5 SYÖTTEEN TARKISTUS- JA VIRHEENKORJAUSMENETELMÄT

11.5.1 Automaattisesti katkaistujen hakusanojen tarkistus

FULLTEXT-projektissa hakusanojen automaattinen katkaisu toteutettiin Finstems- ja Hahmotin-ohjelmilla (testausympäristö T2). Testaajan piti itse hakusanan lisäksi syöttää näille ohjelmille myös hakusanan sanaluokkakoodi. Lisäksi Finstems-ohjelmalle oli syötettävä yhdyssanat siten, että yhdyssanan viimeisen osan eteen lisättiin koodimerkki.

Koska testaaja eli tutkija itse osasi antaa hakusanat oikeassa muodossa ja oikealla sanaluokkakoodilla varustettuna, Finstems ja Hahmotin normaalitapauksessa tuottivat oikeat taivutusvartalot. Todellisissa tuotantojärjestelmissä hakusanojen automaattinen katkaisu kuitenkin tulisi toteuttaa toisin kuin projektin testauksissa, koska hakusanan sanaluokka ei välttämättä ole hakijalle itsestäänselvä.

Jotta hakijalta ei tarvitsisi kysyä hakusanan sanaluokkaa, voitaisiin automaattinen katkaisu toteuttaa esimerkiksi siten, että hakusanan sanaluokaksi oletettaisiin aina substantiivi. Nykyisten hakujärjestelmien käyttäjienhän on todettu hakevan enimmäkseen substantiiveilla. (Luonnollisesti käyttäjälle tulisi antaa mahdollisuus poiketa oletusarvosta niin, että hän voisi halutessaan hakea myös muiden sanaluokkien sanoilla tai toisaalta voisi katkaista hakusanat itse haluamastaan kohdasta.)

Edellä kuvatun oletusarvon ongelmana on se, että ainakin sanomalehtitekstissä adjektiivit ja verbit ovat hyvin käyttökelpoisia hakusanoja, vaikka hakijat eivät niitä huomaisikaan käyttää. Substantiivin asettaminen oletusarvoksi voisi saada käyttäjät tekemään entistä suppeampia hakuja, kun muiden sanaluokkien sanojen käyttöä hakusanoina pitäisi pikemminkin edistää.

Toinen tämän ratkaisumallin epäkohta on, että käyttäjä voikin antaa hakusanaksi jonkin muun sanaluokan sanan kuin substantiivin, muttei huomaa tai osaa antaa sen mukaista sanaluokkakoodia. Kun hakujärjestelmä oletusarvojensa mukaisesti taivuttaa hakusanan substantiivina, tulokseksi voidaan

väärän sanaluokan vuoksi saada liian vähän tai väärin taivutettuja vartaloita. Tätä käyttäjä ei välttämättä havaitse, koska hän "tietää" taivutusvartalo-ohjelmien toimivan suomen kielen sääntöjen mukaisesti. Hakujärjestelmää ei pitäisi rakentaa niin, että hakijan oletetaan aina tarkistavan vartalo-ohjelmien tuottamat hakusanat - automaattisen katkaisun ideanahan nimenomaan on siirtää vastuu hakusanan oikeasta katkaisemisesta hakijalta hakujärjestelmälle.

Edellä esitetyt ongelmat voidaan välttää soveltamalla toista tapaa hakusanan sanaluokan määrittämiseen: annetaan se perusmuoto-ohjelmien (Morfo tai Twol) tehtäväksi. Tällöin hakusana syötetään ensin perusmuoto-ohjelmalle, joka etsii sanakirjastaan oikean sanaluokan. Samalla voitaisiin myös hajottaa yhdyssana osiinsa ja merkitä sen viimeinen osa Finstems-ohjelmaa varten. Tämän jälkeen hakusana ja sanaluokka syötetään taivutusvartalo-ohjelmalle, joka katkaisee hakusanan sopivasta kohdasta.

Näin kuitenkin joudutaan perusmuoto-ohjelmien sanakirjan armoille eli menetetään se etu, että taivutusvartalo-ohjelmat ovat sääntöpohjaisia eivätkä siten tarvitse sanakirjaa tai ylläpitoa. Ne pystyvät taivuttamaan uudissanatkin säännöstönsä perusteella, kunhan sanaluokka on annettu oikein. Jos taas hakusana ei löydy perusmuoto-ohjelman sanakirjasta, ei siellä ole sen sanaluokkaakaan, vaan tämä on pääteltävä muilla keinoin.

Mikäli käyttäjä onkin antanut hakusanan taivutusmuodossa (*Yhtyneet Paperitehtaat*, *Vuoden kylä*), hakusana voidaan ensin palauttaa perusmuotoonsa ja sen jälkeen tuottaa taivutusvartalot tästä perusmuodosta. Näin saadaan ratkaistuksi myös se ongelma, miten taivutusmuodossa annettujen hakusanojen taivutusvartalot tuotetaan oikeista (eli säännönmukaisista) eikä oikeiksi luulluista perusmuodoista.

Toisaalta on mahdollista, että hakusana, jonka perusmuoto-ohjelma tulkitsee sanakirjassaan olevan sanan taivutusmuodoksi, ei olekaan sitä. Jos käyttäjä hakee Salmén-nimeä (ongelmakysely 40) ja se puuttuu sanakirjasta, perusmuoto-ohjelma tulkitsee *salmen*-hakusanan vaikkapa salmi-yleisnimen tai Salme-erisnimen genetiiviksi (edellyttäen, että nämä sanat ovat sanakirjassa). Mikäli tässä tapauksessa haetaan suoraan sanakirjasta löytyvien sanojen vartaloilla, tulokseksi saadaan paljon asiaankuulumattomia dokumentteja, joita ei alkuperäisen hakusanan vartaloilla olisi saatu (salmi, sal-

me, salmea jne). Hakijan on varmasti vaikea käsittää, miksi *salmen*-hakusanalla saadaan tulokseksi dokumentti, jossa esiintyy sana salmia.

Mikäli hakusanaa ei löydetä suoraan perusmuoto-ohjelman sanakirjasta, pitäisi siis ensin selvittää, oliko kyseessä käyttäjän erehdys (taivutusmuoto perusmuodon sijasta) vai sanakirjasta puuttuva sana. Mikäli kyseessä on aidosti tuntematon sana, sen sanaluokka pitäisi määritellä käyttäjän avustuksella, jotta hakusanasta saadaan tuotetuksi oikeat vartalot.

11.5.2 Syötteen tarkistus hakutulosten seulonnan yhteydessä

Tutkimusympäristössä T3 hakusanat ensin katkaistaan automaattisesti ja sitten seulotaan tulokset perusmuoto-ohjelman avulla. Ensimmäiseen vaiheeseen eli hakusanojen automaattiseen katkaisuun liittyvät ongelmat on kuvattu edellisessä luvussa. Seuraava seulontavaihe edellyttää lisäksi, että hakemistosta poimittavien sanojen perusmuodot löytyvät perusmuoto-ohjelman sanakirjasta.

Jos perusmuoto-ohjelma ei voi tunnistaa hakemistosta löytynyttä sanamuotoa, se ei vielä takaa, etteikö hakemistosana voisi olla hakusanan jokin muoto. Jos haetaan katkaistulla hakusanalla *koira**, hakemistosta voi löytyä muun muassa taivutusmuoto koiratrimmaamon. Mikäli tämän yhdyssanan loppuosa trimmaamo puuttuu perusmuoto-ohjelman sanakirjasta, koko yhdyssana jää tunnistamatta, vaikka sen alkuosa koira löytyykin sanakirjasta.

Hakemistosta löytynyt taivutusmuoto voidaan varmasti hylätä ainoastaan silloin, kun se voidaan perusmuotoistaa ja kun perusmuoto todetaan eri sanaksi kuin hakusana. Koska erilaisten heurististen menetelmien tuottama tulos ei välttämättä ole paras mahdollinen, on järkevää soveltaa menetelmiä siten, että epävarmoissa tapauksissa mahdollisesti epärelevantti dokumentti annetaan käyttäjän itsensä arvioitavaksi.

11.5.3 Syötteen tarkistaminen perusmuotohakemistosta haettaessa

Mikäli hakusana on suoraan perusmuoto-ohjelman sanakirjassa oleva sana, sitä voidaan hakea sellaisenaan perusmuotohakemistosta. Ongelmia tulee silloin, kun hakusanaa ei löydy perusmuoto-ohjelman sanakirjasta. Tällöin hakusana on käsiteltävä samalla perusmuoto-ohjelman versiolla ja sanakir-

jalla kuin dokumenttien tekstit hakemistoa rakennettaessa, jotta haku- ja hakemistosanat vastaisivat toisiaan.

Tämä hakusanan tarkistaminen ja muuntaminen hakemistossa käytettyyn muotoon on nimenomaan hakujärjestelmän tehtävä, koska käyttäjän on mahdotonta tietää, miten mikäkin sananmuoto on perusmuotoistettaessa tulkittu. Vaikka käyttäjä tuntisikin perusmuotoja tuottavan ohjelman yleiset toimintaperiaatteet, yksittäisen sananmuodon tulkinta riippuu perusmuotoohjelman säännöistä ja siitä, mitkä sanat sisältyvät sen sanakirjaan.

Perusmuotohakemistojen ongelmana on, että vaikka hakusanan perusmuoto puuttuukin sanakirjasta, voi osa sen taivutusmuodoista olla identtisiä jonkin sanakirjassa olevan sanan perus- tai taivutusmuodon kanssa. Tallennusvaiheessa tällaiset sananmuotohomografit tulkitaan sanakirjassa olevan sanan esiintymiksi, jolloin ne joutuvat tämän väärän perusmuodon yhteyteen perusmuotohakemistoon. Siksi hakuvaiheessa on tutkittava sekä tunnistamattomien sananmuotojen että perusmuotohakemistot, jotta löydetään sekä tuntemattomiksi jääneet että mahdollisesti väärin tulkitut sananmuodot.

Korjaushaku tapahtui syöttämällä hakusana ja sen sanaluokkakoodi Finstems-ohjelmalle, joka tuotti hakusanasta taivutusvartalot (hakusanan sanaluokka on siis saatava selville jollain keinolla; tässä tapauksessa hakijan oli tiedettävä se). Finstems-ohjelman tuottamilla vartaloilla tehtiin kysely molemmista hakemistoista. Lisäksi hakusana syötettiin Morfo-ohjelmalle. Jos Morfo ilmoitti, että hakusana oli sanakirjassa olevan toisen sanan tai toisten sanojen taivutusmuoto, myös nämä perusmuodot haettiin perusmuotohakemistosta. Näin saatua tulosjoukkoa karsittiin edelleen niin, että vartaloilla tehtiin peräkkäishaku itse dokumenttien tekstistä, jolloin jäljelle jäivät vain varsinaisen hakusanan esiintymät ja niiden kanssa homografiset sananmuodot. Yleensä tulokseksi saatiin samat dokumentit kuin hakijan katkaisemilla hakusanoilla. Muutamassa tapauksessa perusmuotohakemiston tulosjoukko sisälsi vähemmän dokumentteja kuin perinteisellä tavalla haettu tulosjoukko, mutta ainoastaan yhdessä tapauksessa (*Sri Lanka*) näin menetettiin relevantteja dokumentteja.

Sri Lanka -tyyppiset virheet voitaisiin periaatteessa välttää siten, että kysely tehdään aina sekä perusmuoto- että tunnistamattomien sananmuotojen hakemistosta. Tällainen ongelmatyyppi ei kuitenkaan liene kovin yleinen, joten näin massiivinen ja resursseja vaativa korjausmenetelmä ei tunnu tarkoituk-

senmukaiselta. Yksi ratkaisumahdollisuus on, että tämäntyyppisistä ongelmasanoista laaditaan poikkeussanojen lista, joka aina käytäisiin läpi ennen varsinaista hakua. Tämä kuitenkin edellyttäisi, että tällaisten sanojen määrä on rajallinen.

11.5.4 Syötteen tarkistaminen kaksoishakemistosta haettaessa

Kaksoishakemistosta haettaessa tulkintavirheitä ei tarvinnut korjata niin monivaiheisesti kuin tutkimusympäristöissä T4 ja T5. Kun hakusanaa ei tutkimusympäristössä T6 löydetty perusmuotohakemistosta, se joko haettiin hakijan itse katkaisemilla hakusanoilla (kuten ympäristössä T1) tai katkaistiin automaattisesti Finstems-ohjelman avulla (kuten tutkimusympäristössä T2). Jälkimmäisessä tapauksessa tulokseksi saatiinkin yleensä samat dokumentit kuin tutkimusympäristössä T2. Poikkeuksena oli *Sri Lanka* -hakusana, jossa Lanka-sanan monitulkintaisuutta ei huomattu perusmuodosta. Tällöin korjaushakua ei tehty, ja kyselyn saanti jäi tutkimusympäristössä T6 alemmaksi kuin tutkimusympäristöissä T1 ja T2.

Tutkimusympäristön T6 etuna on, että nykyisten hakujärjestelmien käyttöön tottuneen hakijan ei välttämättä tarvitse muuttaa toimintatapojaan, vaan hän voi hakea taivutusmuotohakemistosta aivan samalla tavalla kuin ennenkin. Lisäksi hän voi halutessaan käyttää myös perusmuotohakemistoa. Koska taivutusmuotohakemisto sisältää kaikki dokumenteissa esiintyneet taivutusmuodot, saanti ei myöskään voi huonontua homografien väärän perusmuotoistamisen seurauksena, kuten voi käydä perusmuotohakemistoa täydentävässä, pelkästään tuntemattomat sanat sisältävässä hakemistossa. Tosin kaksoishakemistosta haettaessa saanti voi kärsiä muista syistä: esimerkiksi silloin, kun korjaushakua ei huomata tehdä (kuten Sri Lanka).

Toinen kaksoishakemiston ongelma on, miten yhdyssanojen alkuosat haetaan. Jos yhdyssanan alkuosat haetaan pelkästään perusmuotohakemistosta, löydetään vain tunnistettujen sanojen esiintymät. Entä sitten ne yhdyssanat, joita ei ole tunnistettu? Ne ovat tallella taivutusmuotohakemistossa siinä muodossa, jossa ne dokumenttien teksteissä esiintyivät. Jos myös ne halutaan mukaan, tehdään siis kysely taivutusmuotohakemistosta katkaistuilla hakusanoilla. Mutta tällöinhän yhdyssanojen alkuosien haku itse asiassa on sama kuin koko kyselyn suoritus taivutusmuotohakemistosta - eli mihin perusmuotohakemistoa itse asiassa tarvitaan?

Tutkimusympäristöissä T4 ja T5 taivutusmuotoisia hakemistosanoja haettiin katkaistuilla (ja siten epätarkoilla) hakusanoilla vain tunnistamattomien sananmuotojen hakemistosta, joka on perinteiseen taivutusmuotohakemistoon verrattuna vain sen pieni osajoukko. Tällöin hakutulokseksi saadaan pienempi määrä dokumentteja kuin haettaessa koko perinteisestä taivutusmuotohakemistosta. Kaksoishakemistoa käytettäessä hakijan siis pitäisi päättää, haluaako hän mahdollisimman hyvän saannin, jolloin tekee kyselyn taivutusmuotohakemistosta katkaistuilla hakusanoilla, vai haluaako hän tätä paremman tarkkuuden. Jälkimmäisessä tapauksessa hän jättää hakematta taivutusmuotohakemistosta, vaikka tällöin voikin muutama dokumentti jäädä löytymättä. - Vaikka kaksoishakemisto on keino välttää dokumenttien hukkuminen homografien väärintulkinnan vuoksi, voi tietoa siis siinäkin jäädä kadoksiin. Näin hukkaantuvat dokumentit tosin ovat toisia kuin ne, jotka hukkaantuvat homografien väärintulkinnan vuoksi.

Jos hakujärjestelmän rakentaja siis haluaa saada kaikki eri hakuvaihtoehdot tarjolle, tulisi tuottaa kolme eri hakemistoa: ositettu perusmuotohakemisto, tunnistamatta jääneiden sananmuotojen hakemisto ja perinteinen taivutusmuotohakemisto, joka sisältää kaikki dokumenteissa esiintyneet sananmuodot.

Kaksoishakemisto on järkevä myös tapauksissa, joissa aineisto sisältää paljon vieraskielisiä sanoja, joita ei kannata sisällyttää perusmuoto-ohjelman sanakirjaan, tai kun kyselyissä halutaan usein käyttää taivutusmuotoisia hakusanoja (kuten kirjastoluetteloissa etsittäessä kaunokirjallisuutta tai yleensä tietynnimisiä teoksia).

Kaksoishakemiston haittana on, että se vie paljon muistitilaa, enemmän kuin mikään muu projektissa kokeiltu hakemistoratkaisu. Edellä hahmoteltu kolmoishakemisto tietenkin veisi vielä enemmän tilaa kuin kaksoishakemisto. Vaikka se tarjoaakin monipuolisemmat haku- ja virheenkorjausmahdollisuudet kuin mikään tutkimuksessa kokeilluista vaihtoehdoista, on tietysti eri asia, onko se hakujärjestelmän ylläpitäjän kannalta kaupallisesti kannattavaa. Vaikka tallennusmuisti nykyisin on halpaa, suurista tekstimassoista syntyy kuitenkin suuria hakemistoja, joiden käsittely ja varmistus vaativat paljon resursseja.

Jos kaksois- tai kolmoishakemistojen rakentamisen syynä on vain se, että hakutuloksen saannin pelätään huonontuvan homografien väärän tulkinnan

vuoksi, hakemistojen luomiseen ja ylläpitoon tarvittavat kustannukset voivat osoittautua suuriksi suhteessa oletettuun hyötyyn. Koska perusmuotohakemistoon tallennettujen hakemistosanojen alkuperäiset esiintymismuodot ovat löydettävissä dokumenttien varsinaisesta teksteistä, on tavallaan tuhlausta tallentaa sanat hakemistoon sekä perusmuodossa että taivutusmuodossa.

Jos hakujärjestelmästä haetaan enimmäkseen perusmuodoilla, olisi tällaiseen järjestelmään järkevää kehittää menetelmiä, joilla ensin rajataan karkeasti joukko dokumentteja, jotka jokseenkin varmasti sisältävät kaikki relevantit dokumentit ja sitten itse dokumenttien tekstejä hyödyntämällä poimia esiin vain tarpeelliset taivutusmuodot sisältävät dokumentit. Tässä kuvatut hakusanojen tarkistusmenetelmät olivat vasta alustavia kokeiluja, joten jatkotutkimus menetelmien kehittämiseksi olisi tarpeen.

11.6 ONGELMASANOJEN YLEISYYS

Eri hakemistovaihtoehtojen paremmuutta homografien suhteen on hankala määritellä siksi, että näiden tai muun tyyppisten ongelmasanojen yleisyyttä ei tämän tutkimuksen aikaan ollut laskettu. Tämän tutkimuksen kyselyaineisto ei ole riittävä ongelman todellisten mittasuhteiden päättelemiseksi. Jotta tiedettäisiin, millaisia virheitä todellisissa tiedonhauissa esiintyy, tarvittaisiin erillinen tutkimus, jossa kerätään ongelmatapauksia eli perusmuoto-ohjelmalle tuntemattomia sanoja tuotantokäytössä olevasta tunnistamattomien sanojen hakemistosta.

Tällaista tuntemattomien sanojen keräystä ei voi toteuttaa manuaalisesti, ilman tiedonhakujärjestelmää, koska perusmuotoistamisen tulos riippuu monista tekijöistä, kuten perusmuoto-ohjelman säännöistä ja sanastosta. Tutkittava sana saattaa jollain sanastoversiolla olla homografinen ja jollain toisella sanastoversiolla ei, vaikka perusmuotoistamisessa käytetyt kielioppisäännöt olisivatkin samat. Hakujärjestelmästä kerätty ongelmasanojen joukko ei ole absoluuttinen ja muuttumaton joukko. Asiaa ei pidäkään tarkastella yksittäisen sanan tasolla, vaan suhteuttaa ongelmasanojen määrä sanakirjan kokoon ja hakemistosanojen määrään: mikä on ongelmasanojen osuus suhteessa näihin?

Mikäli homografeista yms. johtuvat ongelmat osoittautuisivat paremminkin teoreettisiksi kuin todellisiksi, ei hakujärjestelmiin välttämättä tarvitsisi ra-

kentaa mutkikkaita virheenkorjausmenetelmiä vain muutaman satunnaisen ongelmatapauksen korjaamiseksi, vaan karkeammatkin menetelmät saattavat riittää. Jos hakija painottaisi korkeaa saantiarvoa, rajattaisiin karkeilla menetelmillä suhteellisen suuri ja paljon epärelevantteja dokumentteja sisältävä tulosjoukko, joka annetaan hakijan tarkastettavaksi. Jos taas hakija painottaisi korkeaa tarkkuusarvoa, tulosjoukko lisäksi karsittaisiin automaattisesti menetelmällä, joka mahdollisesti pudottaisi tulosjoukosta joitain relevantteja dokumentteja pois, mutta poistaisi useimmat epärelevantit dokumentit.

Koska perusmuoto-ohjelmat eivät tunnista sananmuotoja, joissa on kirjoitusvirhe, niitä voitaisiin periaatteessa hyödyntää tekstin oikeinkirjoituksen tarkistamiseen. Käytännössä oikeinkirjoituksen tarkistus pitäisi kuitenkin hoitaa jo tekstiä kirjoitettaessa eikä vasta tallennusvaiheessa, jotta siitä saataisiin mahdollisimman suuri hyöty. Vierasperäisten sanojen lisääminen perusmuoto-ohjelman sanakirjaan taas on ongelmallista, koska ne eivät välttämättä käyttäydy suomenkielisten sanojen tavalla. Erisnimet taas, olivatpa ne sitten vierasperäisiä tai suomenkielisiä, ovat alati kasvava sanaryhmä, jonka sisällyttäminen sanakirjaan ei ole edes teoriassa mahdollista.

Hakujärjestelmä onkin rakennettava periaatteella, että sanakirja ei voi olla täydellinen ja että tulkintavirheitä voi tapahtua, joten tulkintavirheet on pystyttävä korjaamaan mahdollisimman automaattisesti. - Tosin kaikkia ongelmia ei tarvitse yrittää ratkoa lingvistiikan keinoin. Jos esimerkiksi tekstien kirjoittajat merkitsevät erisnimet tekstiin valmiiksi jollain sovitulla koodilla, ei perusmuoto-ohjelmiin tarvitse kehittää mitään erityisratkaisuja erisnimien automaattista tunnistusta varten.

11.7 LÄHEISYYSOPERAATTORIN HYÖTY SANALIITTOJA HAETTAESSA

Kun tekstien sanat perusmuotoistetaan, menetetään sanojen taivutusmuotoihin sisältyvää informaatiota. Vuoden kylä on määrätty ilmaus, jossa osien yhdistelmä on muutakin kuin sanojen summa, vrt. "Asuin *vuoden* Syrjäntaan *kylässä*" tai "Hän kävi *kylässä* kaksi *vuotta* sitten". Taivutusmuotohakemistosta on helppo hakea taivutusmuotoisia ilmaisuja. Vastaava tieto löytyy kuitenkin perusmuotohakemistostakin, vaikka mutkan kautta. Ensin haetaan hakusanan perusmuodolla tietty dokumenttijoukko ja sitten käydään näin saatujen dokumenttien tekstit läpi ja poimitaan niistä ne, joissa esiintyy

haluttu hakusanan muoto. Ongelmana kuitenkin on, että jos perusmuotohaku on liian epätarkka, tulosjoukko on suuri ja läpikäytäviä dokumentteja paljon, jolloin hakujärjestelmän vasteaika voi kasvaa käytännössä liian pitkäksi.

Kun hakuavain on sanaliitto, yksi varma keino parantaa kyselyn tarkkuutta on käyttää mahdollisimman tiivistä läheisyysoperaattoria. Kun testikyselyissä sanaliiton eri osat kytkettiin toisiinsa virkeoperaattorilla, löydettiin kaikki relevantit dokumentit. Kun nämä samat sanaliiton sanat yhdistettiin toisiinsa JA-operaattorilla virkeoperaattorin sijasta, ei tulosjoukkoon tullut uusia relevantteja dokumentteja. Sen sijaan tulosjoukon tarkkuus huononi, kun liian väljän operaattorin seurauksena tulosjoukkoon kertyi lisää epärelevantteja dokumentteja. Vakiokyselyissäkin todettiin, että pelkästään virkeoperaattorin käyttö nosti tulosjoukon tarkkuusarvon useimmiten jo niin korkeaksi, että kyselytyypillä ei enää ollut vaikutusta tarkkuusarvoon - kyselytyyppien väliset erot eivät olleet tilastollisesti merkitseviä eikä niillä siten ollut käytännön merkitystä.

Tutkimuksessa käytetyssä BASIS-K-järjestelmässä tarkin mahdollinen osoite oli virke. Mikäli hakusanojen keskinäinen sijainti olisi ollut mahdollista ilmaista tätäkin täsmällisemmin, esimerkiksi määritellä hakusanojen sijaitsevan välittömästi peräkkäin, hakujen tarkkuusarvot olisivat olleet vielä korkeammat.

Hakujärjestelmiä kehitettäessä onkin siis syytä ottaa huomioon, että kaikkiin mahdollisiin ongelmiin ei ole aina tarkoituksenmukaista etsiä lingvistisiä ratkaisuja, vaan monesti hakujärjestelmän perinteiset keinot, kuten läheisyysoperaattorit, voivat ratkaista asian helpoiten. Vakiintuneiden hakujärjestelmätekniikoiden, heuristiikkojen ja lingvististen keinojen sopivalla yhdistämisellä voidaan käytännössä yltää käyttäjiä tyydyttävään eli kaupallisesti riittävään tasoon.

12 JOHTOPÄÄTÖKSET

Tässä tutkimuksessa on selvitetty, miten suomen kielen morfologisten tulkinta-ohjelmien avulla voidaan ratkaista joitakin sellaisia tiedon tallennuksen ja haun ongelmia, jotka johtuvat suomen kielen erityispiirteistä. Tutkimus oli luonteeltaan laboratorioympäristössä toteutettu evaluointitutkimus, jota varten laadittiin oma kyselykokoelma ja joukko vertailtavia hakemistoratkaisuja ja tutkimusympäristöjä.

Miten dokumentteja sitten haetaan? Periaatteessa siten, että tiedontarvitsija määrittelee, millaisia ominaisuuksia dokumentilla tulisi olla, jotta se olisi hänen tiedontarpeensa kannalta relevantti. Perinteisissä hakujärjestelmissä tämä tiedontarpeen määrittely tapahtuu käytännössä hyvin matalalla abstraktiotasolla: dokumentteja haetaan sillä perusteella, millaisia merkkijonoja - luonnollisen kielen tapauksessa sananmuotoja - dokumenttien tekstissä on esiintynyt. Koska sanasta (lekseemistä) voi olla monia erilaisia ilmentymiä, hakijan on arvioitava, millaisissa eri taivutusmuodoissa eli millaisina merkkijonovakioina hakusana on esiintynyt. Vasta tämän päättelyn perusteella hän voi muodostaa katkaistun hakusanan tai hakusanat eli merkkijonokaavion tai -kaaviot, jotka kattavat kaikki eri merkkijonovakiot.

Tässä tutkimuksessa selviteltiin periaatteita, joilla abstraktiotasoa voidaan nostaa merkkijonotasolta sanatasolle: hakijan ei tarvitsisi hallita suomen kielen sanojen taipumista yms. merkkijonotason seikkoja, vaan pelkkä hakusanan keksiminen riittää. Hakujärjestelmän tehtävänä on huolehtia, että kaikki eri muodoissa esiintyneet hakusanan esiintymät tulisivat löydetyksi.

Abstraktiotason nostaminen merkkijonoista sanoiksi ei tietenkään riitä ratkaisemaan kuin osan tiedonhaun ongelmista. Ensinnäkin tarvitaan keinot myös niiden merkkijonojen hakemiseen, jotka eivät ole luonnollisen kielen sanoja. Toiseksi, sanataso on edelleenkin liian matala abstraktiotaso. Tiedontarvitsijan ja tekstin tuottajan sanavalinnat eivät välttämättä ole identtisiä: kaikissa relevanteissa dokumenteissa ei käytetä juuri niitä sanoja, jotka kyselyssä esiintyvät (Wormell 1984; Blair & Maron 1990). Sitä paitsi kaikilla tieteenaloillakaan ei ole yleistä yksimielisyyttä siitä, mitä tietyt sanat tarkoittavat, vaan eri kirjoittajat voivat käyttää samoja sanoja eri merkityksissä (Harter 1986, s. 34 - 35).

Kuva 14. Tiedonhaun tasoperiaate

Sanatasolta on siis astuttava askel ylemmäksi eli käsitetasolle (kuva 14). Silloin lähtökohtana on, että hakija ensin määrittelee tiedontarvetta kuvaavat käsitteet ja sen jälkeen nämä käsitteet kielellistetään esimerkiksi luonnollisen kielen sanoiksi. Ideaalihakujärjestelmässä hakijan ei tarvitsisi tehdä kuin käsitetason määrittelyt ja hakujärjestelmä huolehtisi käsitteiden muuntamisesta ilmaisutasolle (Järvelin et al. 1996; Kekäläinen 1999).

Kun hakemiston sanat ovat perusmuodossa, tesaurusten, sanakirjojen ja hakemistojen linkittäminen yhteen helpottuu olennaisesti tai hakemistoja voidaan hyödyntää tesaurusten tuottamisessa (Thönssen 1988). Myös sanojen esiintymistiheyteen perustuvien relevanssikertomien[1] laskenta on yksinkertaisempaa perusmuotohakemistossa. Taivutusmuotohakemistossa ei hakemistosanojen taajuuksia ei kannata suoraan laskea, koska tällöin laskettaisiin saman sanan eri esiintymät erikseen. Kun taivutusmuodot normalisoidaan perusmuodoiksi, lasketaan sanojen eikä taivutusmuotojen esiintymistiheyksiä.

Tutkimuksessa todettiin, että kun kyselyä laajennetaan hakusanan johdosperheellä, hakusanan sisältävillä yhdyssanoilla tai yhdyssanan osilla, hakutuloksen saanti paranee. Tämähän ei ole mitenkään yllättävää, vaan täysin järkeenkäypä tulos: mitä useampia verkkoja veteen heittää, sitä todennäköisemmin johonkin niistä osuu saalista. Sen sijaan kiinnostavaa on tutkia, mitä samanaikaisesti tapahtuu hakutuloksen tarkkuudelle. Tässä suhteessa eri tutkimusympäristöt ja kyselytyypit käyttäytyivät eri tavoin.

Perusmuotohakemistosta saadaan tarkempia tuloksia kuin taivutusmuotohakemistosta: kun molemmissa käytetään täsmälleen samalla tavalla katkaistuja hakusanoja, perusmuotohakemistosta saatujen tulosjoukkojen tarkkuus on parempi kuin samanlaisella kyselyllä taivutusmuotohakemistosta saatujen tulos-

[1] Osittaistäsmäytykseen perustuvissa hakujärjestelmissä voidaan relevanssikertoimien avulla määritellä haku- tai hakemistosanalle halutunlainen painoarvo. Kertoimien avulla voidaan esimerkiksi määritellä, että tekstissä harvoin esiintyvä sana on tärkeämpi ja siten saa suuremman kertoimen kuin jokin toinen, usein esiintyvä sana. Kun tulosjoukon dokumentit järjestetään niiden relevanssiarvon mukaiseen järjestykseen, tulevat ensimmäiseksi dokumentit, joissa esiintyvillä sanoilla on korkeimmat painoarvot.

joukkojen tarkkuus. Tämän perusteella kyselyn laajentaminen siis olisi perusmuotohakemistossa "riskittömämpää" kuin taivutusmuotohakemistossa.

Toisaalta kävi selvästi ilmi, että kyselyissä ei kannata käyttää pelkkiä hakusanan perusmuotoja, vaikka tällaisten kyselyjen tarkkuus olisikin hyvä. Tällaisten kyselyjen saanti nimittäin on huono - kun kyselyä laajennetaan johdosperheellä tai yhdyssanoilla tai näiden osilla, saanti paranee useampia prosenttiyksikköjä enemmän kuin toisaalta tarkkuus prosenttiyksikköinä laskien huononee.

Yleensäkin hakijan on perusmuotohakemistoista hakiessaan kiinnitettävä enemmän huomiota siihen, että myös johdokset ja yhdyssanat on otettu kyselyssä huomioon - kun haetaan taivutusmuotohakemistosta hakijan katkaisemilla hakusanoilla, johdokset ja yhdyssanat usein tulevat ikään kuin kylkiäisinä. Perinteistä hakutapaa ei siis perusmuotohakemistossa voida korvata niin, että hakuvaiheessa annetaan muuten samat hakusanat kuin perinteisestikin haettaessa, mutta vain jätetään ne katkaisematta. Tämä ehkä vaatii hakijalta tietoisempia valintoja kuin perinteisellä tavalla haettaessa, mutta toisaalta antaa hänelle enemmän valinnanvaraa päättää, painottaako kyselyssä saantia vai tarkkuutta. Perinteisessä hakutavassa tällaiseen hienosäätöön ei ole vastaavia mahdollisuuksia.

Tarkkuuden suhteen voitiin todeta myös, että kun hakusanat oli kytketty toisiinsa virkeoperaattorilla, eri kyselytyypeillä saatujen tulosjoukkojen tarkkuusarvojen väliset erot jäivät pieniksi. Tämän tutkimuksen suppeiden vertailujen perusteella ei voi tehdä erityisiä päätelmiä operaattorien merkityksestä hakutulokseen (kattava tutkimus aiheesta: ks. Sormunen 2000), mutta se kuitenkin voidaan todeta, että operaattorien osaavalla hyödyntämisellä voidaan jäljittää esimerkiksi perusmuotoistamisen seurauksena hämärtyneet sanaliitot (kuten Kansan Uutiset -> kansa uutinen). Eli tiedonhaun ongelmiin ei kannata etsiä puhtaasti lingvistisiä ratkaisuja, vaan optimaalista ratkaisua tulisi etsiä yhdistämällä lingvistiset keinot tiedonhakujärjestelmien perinteisiin apukeinoihin.

Tässä tutkimuksessa tutkittiin johdosten ja yhdyssanojen käyttäytymistä erilaisissa tiedonhakutilanteissa. Johdokset ja yhdyssanat olivat tutkimuskohteina siitä syystä, että ne ovat morfologisen tason ilmiöitä ja niiden käsittelyä voidaan tietyssä määrin automatisoida. On täysin mahdollista, että joidenkin toisentyyppisten luonnollisen kielen ilmausten, kuten synonyymien tai yleensäkin semanttisen tason ilmiöiden osuus tiedonhaun tuloksiin on merkittävämpi kuin näiden

tutkittujen kahden morfologian osa-alueen. FULLTEXT-projektissa hyödynnetyillä morfologisilla eli sananmuotoja käsittelevillä ohjelmilla ei pystytä ratkaisemaan sanojen merkitykseen liittyviä ongelmia. Niiden avulla voidaan kuitenkin luoda paremmat edellytykset tällaisten ongelmien ratkaisuun tähtääville menetelmille, kuten edellämainituille hakutesauruksille, sanakirjoille ja relevanssikertoimien laskennalle.

Vaikka haku- ja hakemistosanojen perusmuotoistaminen tekeekin tiedonhaun yksinkertaisemmaksi verrattuna siihen, että hakija joutuu katkaisemaan hakusanat, perusmuoto-ohjelman sanakirjan kattavuus on yksi riskitekijä. Jos jokin sana puuttuu sanakirjasta, seurauksena voi olla virhetulkintoja, joiden korjaaminen hakuvaiheessa on, jos ei mahdotonta, niin ainakin kallista. Käyttäjän kannalta hakujärjestelmä ei tietenkään saisi hukata oleellista tietoa, joten tarvitaan menetelmiä, joilla morfologisten tulkintaohjelmien tekemät virhetulkinnat korjataan automaattisesti. Toisaalta tässä törmätään kustannuksiin: vaikka virheettömyys teoriassa onkin tärkeää, niin kuinka paljon se käytännössä saa maksaa? Kannattaako hakujärjestelmään rakentaa mutkikkaat ja resursseja kuluttavat virheenkorjausmenetelmät, jos sen käyttäjille yleensä riittää vähempi kuin täydellinen sadan prosentin saanti ja jos käyttäjät eivät halua maksaa täydellisen saannin varmistamisesta koituvia kustannuksia - etenkin kun samalla käytännössä joudutaan tinkimään tarkkuudesta.

Sitä paitsi täydellistä hakemistoratkaisua ei ole tarjolla. Kun haetaan hakemistosta, joka ei sisällä yhdyssanojen keski- ja loppuosia omina hakemistosanoinaan, käytännössä menetetään yhdyssanojen keski- ja loppuosina esiintyneet sanat. Kun taas haetaan perusmuodoilla, jäävät virheelliset perusmuototulkinnat, lyhennetyt sananmuodot (kuten Yhtyneet Paperiteht.) yms. löytymättä.

Jos haluttaisiin tarjota mahdollisimman monipuoliset hakumahdollisuudet, pitäisi tietokannan dokumenteista itse asiassa tuottaa kolme eri hakemistoa (ja lisäksi muut, muiden kuin lingvististen syiden perusteella tuotettavat hakemistot):

- kaikki sananmuodot sellaisinaan sisältävä hakemisto (taivutusmuotohakemisto)
- ositettu perusmuotohakemisto
- tunnistamattomat sananmuodot sisältävä hakemisto

Eri tilanteissa sitten valittaisiin kyseiseen tilanteeseen sopiva hakemisto tai yhdistettäisiin haku useasta hakemistosta. Tällainen kolmoishakemisto kuitenkin tarvitsisi paljon muistitilaa. Tosin tämä ei nykyisin tuottane erityisiä ongelmia, onhan tallennusmuistin hinta huomattavasti alempi kuin tiedonhakujärjestelmien kehityksen alkuvaiheessa.

Informaatiotutkija ei välttämättä pidä (kielitieteellisiä) virhetulkintoja niin kriittisenä ongelmana kuin kielitieteilijä, koska hänelle tiedonhaun suhteellisuus tai satunnaisuus on joka tapauksessa tosiasia ("inherently probabilistic nature of the information retrieval process" - Doszkocs 1986). Vaikka kyselyssä oleva yksittäinen sana saataisiinkin käsiteltyä kaikin puolin kielitieteellisesti virheettömästi, perusongelmana on, miten hyvin tämä sana kuvaa tiedontarvetta. Osaako tiedontarvitsija kielellistää tiedontarpeensa oikein? Yksittäinen sana on kyselyssä vain yksi osatekijä - hakijan on osattava hahmottaa kokonaisuus ja otettava huomioon hakuavaimen itsensä lisäksi myös sen suhteet muihin hakuavaimiin.

Seuraavassa luetellaan muutamia FULLTEXT-projektissa esiin nousseita tutkimusaiheita:

Ensimmäinen, olennaisen tärkeä asia on edustavan testikokoelman (dokumenttien, kyselyjen ja relevanssiarvioiden) rakentaminen. Testeissä käytetyn aineiston luonne vaikuttaa hakutuloksiin. On eri asia tehdä testihaut viitetietokannasta ja kokotekstitietokannasta. Krovetzin (1993) ja Hullin (1996) tutkimusten mukaan dokumenttien ja kyselyjen pituudet vaikuttavat siihen, miten suuri hyöty pääteainesten karsinnasta on. FULLTEXT-tutkimuksessa ei käytetty erilaisia dokumenttityyppejä, joten tämän tutkimuksen perusteella ei voida arvioida, mikä vaikutus dokumenttien pituudella on hakutulokseen.

Vaikka tässä tutkimuksessa käytetty aineisto olikin edustava verrattuna alalla tuolloin käytettyihin testikokoelmiin, olisi mielenkiintoista tutkia eri osa-alueita (johdoshaut, yhdyssanahaut) tarkemmin suuremmalla kyselyaineistolla. On kuitenkin selvää, että tällaisen testikokoelman luominen ja täydentäminen ei ole yksittäisen tutkijan puuhastelua, vaan vaatii paljon resursseja. Olisikin tärkeää, että esimerkiksi Tampereen yliopiston informaatiotutkimuksen laitoksen tiedonhakulaboratoriota edelleen laajennetaan ja täydennetään. Vaikka suomalaisen testikokoelman kehitystyöhön ei voidakaan panostaa kuin murto-osa siitä mitä tarvitaan esimerkiksi kansainväliseen TREC-hankkeeseen, on kansallisen

tiedonhakulaboratorion kehittäminen tärkeää: ulkomaiset tahot eivät suomenkielistä aineistoa kokoa eivätkä tutki.

Toinen testijärjestelyihin liittyvä asia on ongelmailmausten taajuuden selvittäminen, jotta tiedettäisiin, minkä kokoluokan ongelmasta on kyse. Perusmuotoistamisesta aiheutuvat ongelmatapaukset pitäisi laskea ja tyypitellä systemaattisesti laajan empiirisen aineiston perusteella. Aineisto kerättäisiin todellisen, tuotantokäytössä olevan perusmuotohakemiston loki- eli seurantatiedostosta. Näin saataisiin tietoa siitä, millaisilla ongelmatyypeillä todella on vaikutusta tiedonhaun tuloksiin. Merkittäviksi todetuille ongelmailmauksille kehitetään sopiva virheenkorjausmenetelmä, mikäli todetaan, että karkeat korjausmenetelmät eivät riitä.

FULLTEXT-projektissa ei hyödynnetty ohjelmia, jotka olisivat disambiguoineet eli yksiselitteistäneet morfologisen analyysin tuloksia, koska tällaisia ohjelmia ei projektin aikana vielä ollut saatavilla suomen kieltä varten. Miksi morfologisen analyysin sitten yleensä sallitaan tuottavan monitulkintaisuuksia? Perustana on se näkemys, että kieliopillisen analyysin halutaan olevan varmasti oikea. Jos jollekin ilmaukselle halutaan ehdottomasti löytää vain yksi ainoa luenta, vaarana on, että se onkin väärä. Monitulkintaisuus katsotaan pienemmäksi ongelmaksi kuin virheellinen tulkinta. Koska disambiguointimenetelmiä on FULLTEXT-projektin jälkeen kehitetty huomattavasti (Voutilainen 1994; Conexor 2000), ei niiden muokkaamisessa ja liittämisessä toimivaan tiedonhakujärjestelmään pitäisi olla teoreettisia ongelmia.

Homografeista on muissa tutkimuksissa todettu, että ne tuottavat vain vähän ongelmia käytännön tiedonhaussa (Wormell 1984; Lancaster 1986, s. 69, Leppänen 1996). Käytännössä kyselyt tavallisesti sisältävät useita konjunktiivisia hakuavaimia, jotka riittävät rajaamaan tulosjoukosta pois dokumentit, joissa homografi on väärän sanan esiintymä. Perusmuotohakemistoissa väärintulkitut homografit voivat huonontaa tarkkuutta siksi, että tulosjoukkoon päätyy perusmuotoistamisen vuoksi myös sellaisia hakemistosanan taivutusmuotoja jotka alunperin eivät olleet hakusanan kanssa homografisia. Tällaiset väärintulkinnan vuoksi tulosjoukkoon päätyneet dokumentit on kuitenkin mahdollista karsia pois automaattisesti (luku 10.3).

Perusmuotohakemistossa yksiselitteistämistä tarvitaan lähinnä hakemiston siistimiseksi (jotta vaikkapa keskusohjaimen-muodosta ei joutuisi hakemistoon tul-

kintaa keskusohjatimeä). Ylimääräisten tulkintojen poistaminen hakemistosta säästää muistitilaa ja tekee hakemistosta käyttäjän kannalta järkevämmän, kun sieltä jäävät pois paperiarkinen-hakemistosanan kaltaiset oudot ilmaukset. Lisäksi yksiselitteistäminen parantaa hakujen tarkkuutta (esimerkiksi jottei *imeä*-hakusanalla saada tulokseksi myös muotoa keskusohjatimeä).

FULLTEXT-projektissa käytetyt perusmuoto-ohjelmat poistivat vain sijapäätteet eivätkä koskeneet johtimiin. Näistä ohjelmista on kuitenkin mahdollista rakentaa myös johtimia käsittelevät versiot. Silloin voitaisiin tutkia, voidaanko hakusanasta tuottaa automaattisesti sekä sen kantasana että kaikki sen johdokset. Tällöin voisi teoriassa olla mahdollista lisätä hakusanan johdosperhe kyselyyn pelkästään lingvistisin keinoin, ilman hakutesauruksen tms. apua. Ongelmaksi voi kuitenkin tulla yligenerointi: täysin sääntöpohjaisesti toimiva ohjelma tuottaa myös johdoksia, joita todellisuudessa ei ole. Tämä voitaisiin kiertää tekemällä tarkistuskierros, jossa ensin tutkitaan hakemistosta, löytyykö kyseinen johdos sieltä. Hakijalle voitaisiin sitten näyttää vain ne johdokset, joiden olemassaolo on hakemiston perusteella varmistettu. Toinen mahdollisuus on hyödyntää johdosten generointia hakutesauruksen laatimisessa. Joka tapauksessa lähteenä käytettävä hakemisto on pitänyt tuottaa riittävän suuresta tekstimassasta (kymmeniä tuhansia lehtiartikkeleita), jotta sen voi katsoa sisältävän edustavan määrän eri johdoksia.

Sulkusanalistan hyötyjä suomenkielisen tekstin tallennuksessa ja haussa on pidetty vähäisinä. Joissain suomalaisissa sovelluksissa sulkusanalista on rakennettu, toisissa taas ei. Sulkusanalistan vaikutusta tiedonhaun tuloksiin tulisi selvittää vertailututkimuksen avulla.

Sanaliitot ja nimet käyttäytyvät usein poikkeavasti, esimerkiksi sanaliiton alkuosa ei aina taivu jälkiosan kanssa samalla tavalla (Suomen Pankki : Suomen Pankissa). Tiedonhaku helpottuisi, jos sanaliitot ja nimet voitaisiin tunnistaa tekstistä automaattisesti. Ainakin englannin kielessä nimien automaattista tunnistamista on jo tutkittu: Raun (1991) tutkimusryhmä kehitti heuristisesti toimivan tekoälyohjelman, joka poimi talousuutisista yritysten nimiä. Kun ohjelman tuotosta verrattiin verrattiin ihmisindeksoijan tuottamiin nimiehdotelmiin, ohjelma löysi noin 95 prosenttia nimistä, joita ihmisindeksoija oli esittänyt. Tämän lisäksi ohjelma löysi uusia, aivan kelvollisia nimiä, joita ihminen ei ollut ehdottanut - tällaisia lisälöytöjä oli kaikkiaan 40 prosenttia ohjelman nimiehdotelmista. Vastaavia tutkimuksia on ollut myös Suomessa (esimerkiksi Lah-

tinen 1995). Toisaalta nimien käsittelyä voidaan helpottaa muillakin kuin kielitieteellisillä keinoilla. Ne voitaisiin esimerkiksi koodata tekstiin mukaan jo kirjoitusvaiheessa siten, että kirjoittajat itse merkitsisivät, mitkä tekstikatkelmat ovat erisnimiä.

Vaikka Harman (1991) epäilikin, ettei pääteainesten karsinnan hyödyllisyyttä ole riittävästi osoitettu, ovat monet muut tutkijat (Keen 1991; Lennon 1981; Krovetz 1993; Hull 1996) kuitenkin todenneet, että karsinnaista on hyötyä myös englannissa. Morfologisesti mutkikkaammassa kielessä, kuten suomessa, perusmuotojen käyttämisen hyödyt on helpompi todeta. Yksi mielenkiintoinen tutkimusalue olisi tutkia suomen kielellä toimivaksi todettuja menetelmiä muunkielisissä tietokannoissa ja toisaalta kielten välisessä tiedonhaussa (vrt. Pirkola 1999). On mahdollista, että suomen kielen ongelmien ratkaisuun kehitetyt menetelmät ovat hyödyllisiä myös muissa kielissä. Yhtä hyvin voi osoittautua, että jotkin menetelmät toimivat suomen kielessä hyvin, mutta huonosti toisenrakenteisissa kielissä. Kummassakin tapauksessa tietämyksemme luonnollisten kielten rakenteista ja niiden vaikutuksesta tiedonhakuun paranisi.

Edellä on esitetty joukko kieliteknologiaan liittyviä tiedonhaun tutkimusaiheita. Olennaista on, että perusmuotohakemistojen myötä suomenkielistä aineistoa käsittelevät hakujärjestelmät pääsevät kuvaannollisesti sanoen samalle lähtöviivalle englanninkielistä aineistoa käsittelevien ja englannin kielen ehdoilla toimivien hakujärjestelmien kanssa. Näin suomenkielisen tekstin tallennuksessa ja haussa voidaan aiempaa helpommin ottaa muualla syntyneitä tutkimus- ja kehitysideoita käyttöön.

LÄHDELUETTELO

PAINETUT LÄHTEET

Abu-Salem, H., Al-Omari, M. & Evens, M. W. 1999. Stemming methodologies over individual query words for an Arabic information retrieval system. Journal of the American Society for Information Science, vol. 50, nro 6, s. 524 - 529.

Alkula, R. & Honkela, T. 1992. Tekstin tallennus- ja hakumenetelmien kehittäminen suomen kielen tulkintaohjelmien avulla: FULLTEXT-projektin loppuraportti. Espoo: Valtion teknillinen tutkimuskeskus. (VTT Julkaisuja 765.) 104 s. + liitt. 18 s. ISBN 951-38-4113-8

Arampatzis, A. T., Tsoris, T., Koster, C. H. A. & van der Weide, Th. P. 1998. Phase-based information retrieval. Information Processing and Management, vol. 34, nro 6, s. 693 - 707.

Arppe, A. 1996. Information explosition and the use of linguistic tools in Finland. Teoksessa: Harakka, T. & Koskela, M. (toim.) Kieli ja tietokone. AFinLAn vuosikirja 1996. Jyväskylä: Suomen soveltavan kielitieteen yhdistys. (Suomen soveltavan kielitieteen yhdistyksen AFinLA julkaisuja 54.) S. 7 - 32. ISBN 951-9388-42-7

Atk-sanakirja 1990. 5. p. Espoo: Suomen Atk-kustannus. 268 s. ISBN 951-762-152-3.

Bain, M. et al. 1989. Free text retrieval systems: a review and evaluation. London, Taylor Graham. 120 s. 0-947568-42-5

Beaulieu, M., Robertson, S. & Rasmussen, E. 1996. Evaluating interactive systems in TREC. Journal of the American Society for Information Science, vol. 47, nro 1, s. 85 - 94.

Belkin, N. J. 1984. Cognitive models and information transfer. Social Science Information Studies, vol. 4, nro 2 & 3, s. 111 - 129.

Bell, C. & Jones, K. P. 1979. Towards everyday language information retrieval systems via minicomputers. Journal of the American Society for Information Science, vol. 30, nro 6, s. 334 - 339.

Blair, D. C. 1990. Language and representation in information retrieval. Amsterdam: Elsevier. 335 s. ISBN 0-444-88437-8

Blair, D. C. & Maron, M. E. 1990. Full-text information retrieval: further analysis and clarification. Information Processing and Management, vol. 26, nro 3, s. 437 - 447.

Borlund, P. 2000. Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation, vol. 56, nro 1, s. 71 - 90.

Borlund, P. & Ingwersen, P. 1997. The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation, vol. 53, nro 3, s. 225 - 250.

Brooks, H. M. & Belkin, N. J. 1983. Using discourse analysis for the design of information retrieval interaction mechanisms. ACM Sigir Forum, vol. 17, nro 4, s. 31 - 47.

Brooks, H. M., Daniels, P. J. & Belkin, N. J. 1985. Problem descriptions and user models: developing an intelligent interface for document retrieval systems. Teoksessa: Informatics 8: Advances in intelligent retrieval. Proceedings of a conference at Wadham College, Oxford, 16 - 17 April 1985. London: Aslib, s. 191 - 214. ISBN 0-85142-195-4

Clemencin, G. 1988. Querying the French Yellow Pages: Natural language access to the directory. Information Processing and Management, vol. 24, nro 6, s. 633 - 649.

Conover, W. J. 1980. Practical nonparametric statistics. 2. ed. New York: Wiley. 493 s. ISBN 0-471-02867-3

Cosijn, E. & Ingwersen, P. 2000. Dimensions of relevance. Information Processing and Management, vol. 36, nro 4, s. 533 - 550.

Crestani, F. 1999. Vocal access to a newspaper archive: Design issues and preliminary investigations. Teoksessa: Fox, E. A. & Rowe, N. (toim.) Digital Libraries '99. The Fourth ACM Conference on Digital Libraries. Berkeley, 11 - 14 August 1999. New York: The Association for Computing Machinery. S. 59 - 66. ISBN 1-58113-145-3

Croft, W. B., Turtle, H. R. & Lewis, S. S. 1991. The use of phrases and structured queries in information retrieval. Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. Chicago, 13 - 16 October 1991. New York: The Association for Computing Machinery. S. 32 - 45. ISBN 0- 89791-448-1

Daniels, P. 1986. The user modelling function of an intelligent interface for document retrieval systems. Teoksessa: Brookes, B. C. (toim.) Intelligent information systems for the information society. Proceedings of the sixth international research forum in information science (IRFIS 6). Frascati, 16 - 18 September 1985. Amsterdam: Elsevier (North -Holland). S. 162 - 176.

Das-Gupta, P. 1987. Boolean interpretation of conjunctions for document retrieval. Journal of the American Society for Information Science, vol. 38, nro 4, s. 245 - 254.

Davies, M. 1991. The implementation of BASIS at the Imperial Cancer Research Fund. Program, vol. 25, nro 3, s. 187 - 206.

Doszkocs, T. E. 1983. CITE NLM: Natural-language searching in an online catalog. Information Technology and Libraries, vol. 2, nro 4, s. 364 - 380.

Doszkocs, T. E. 1986. Natural language processing in information retrieval. Journal of the American Society for Information Science, vol. 37, nro 4, s. 191 - 196.

Erman, L. D., Hayes-Roth, F., Lesser, V. R. & Reddy, D. R. 1980. The Hearsay-II speech-understanding system: intergrating knowledge to resolve uncertainty. Computing Surveys, vol. 12, nro 2, s. 213 - 253.

Fagan, J. L. 1989. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. Journal of the American Society for Information Science, vol. 40, nro 2, s. 115 - 132.

Fidel, R. 1987. Database design for information retrieval: A conceptual approach. New York: John Wiley. 232 s. ISBN 0-471-82786-X

Froehlich, T. J. 1994. Relevance reconsidered - Towards an agenda for the 21st century: Introduction to special topic issue on relevance research. Journal of the American Society for Information Science, vol. 45, nro 3, s. 124 - 134.

Gibb, F. & Smart, G. 1990. Structured information management using new techniques for processing text. Online Review, vol. 14, nro 3, s. 159 - 171.

Glassco, R. A. 1993. Evaluating commercial text search-and-retrieval packages. Information Technology and Libraries, vol. 12, nro 4, s. 413 - 421.

Glavitsch, U. & Schäuble, P. 1992. A system for retrieving speech documents. Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, June 21 - 24, 1992. New York: The Association for Computing Machinery. S. 168 - 176. ISBN 0-89791-524-0

Green, R. 1991. The profession's models of information: A cognitive linguistic analysis. Journal of Documentation, vol. 47, nro 2, s. 130 - 148.

Hakulinen, A. & Ojanen, J. 1976. Kielitieteen ja fonetiikan termistöä. Helsinki: Suomalaisen Kirjallisuuden Seura. (Toimituksia 324.) 170 s. ISBN 951-717-096-3

Harman, D. 1987. A failure analysis on the limitations of suffixing in an online environment. Proceedings of the Tenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. New Orleans: 3 - 5 June 1987. New York: The Association for Computing Machinery. S. 102 - 108. ISBN 0-89791-232-2

Harman, D. 1988. Towards interactive query expansion. Proceedings of the 11th International Conference on Research and Development in Information Retrieval. Grenoble: 13 - 15 June 1988. Grenoble: Presses Universitaires de Grenoble. S. 321 - 331. ISBN 0-89791-274-8

Harman, D. 1991. How effective is suffixing? Journal of the American Society for Information Science, vol. 42, nro 1, s. 7 - 15.

Harman, D. 1993. Overview of the first TREC conference. Proceedings of the Sixteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA: 27 June - 1 July 1993. S. 36 - 47. ISBN 0-89791-605-0

Harter, S. P. 1986. Online information retrieval: concepts, principles, and techniques. San Diego: Academic Press. 259 s. ISBN 0-12-328455-4

Harter, S. P. 1992. Psychological relevance and information science. Journal of the American Society for Information Science, vol. 43, nro 9, s. 602 - 615.

Hattery, M. 1993. Natural language searching. Information Retrieval & Library Automation, vol. 29, nro 6, s. 2 -3.

Honkela, T. 1997. Self-organizing maps in natural language processing. Espoo: Helsinki University of Technology. Väitöskirjan WWW-versio, osoitteessa: <URL: http://www.cis.hut.fi/~tho/thesis/index.html>.

Honkela, T., Kaski, S., Lagus, K. & Kohonen, T. 1996. Newsgroup exploration with WEBSOM method and browsing interface. Espoo: Helsinki University of Technology. (Report A32.) 13 s. ISBN 951-22-2949-8

Hull, D. 1993. Using statistical testing in the evaluation of retrieval experiments. Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA 27 June - 1 July 1993. New York: The Association for Computing Machinery. S. 329 - 338. ISBN 0-89791-605-0

Hull, D. 1996. Stemming algorithms: a case study for detailed evaluation. Journal of the American Society for Information Science, vol. 47, nro 1, s. 70 - 84.

Ingwersen, P. 1992. Information retrieval interaction. London: Taylor Graham. 246 s. ISBN 0-947568-54-9

Ingwersen, P. 1996. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. Journal of Documentation, vol. 52, nro 1, s. 3 - 50.

Jones, K. P. & Bell, C. L. M. 1986. MORPHS - an intelligent retrieval system. ASLIB Proceedings, vol. 38, nro 3, s. 71 - 79.

Jäppinen, H. et al. 1983. Knowledge engineering approach to morphological analysis. Proceedings of the First Conference of the European Chapter of the Association for Computational Linguistics, Pisa. Teoksessa: Jäppinen, H. et al. Morphological analysis of Finnish word forms. Selected reprints. Helsinki: SITRA. (Publications of the Kielikone-project, Series A 1.) S. 49 - 51.

Järvelin, K. 1995. Tekstitiedonhaku tietokannoista: Johdatus periaatteisiin ja menetelmiin. Espoo: Suomen ATK-kustannus. 273 s. ISBN 951-762-297-X

Järvelin, K. & Niemi, T. 1990. Hajautettujen faktatietokantojen käytön yksinkertaistaminen. Kirjastotiede ja informatiikka, vol. 9, nro 1, s. 3 - 16.

Järvelin, K., Kristensen, J., Niemi, T., Sormunen, E. & Keskustalo, H. 1996. A deductive data model for query expansion. Proceedings of the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. Zürich, 18. - 22. August 1996. New York: The Association for Computing Machinery. S. 235 - 249. ISBN 0-89791-714-6

Järvinen, P. & Kerola, P. 1981. Systemointi I: Käytäntö tietosysteemin rakentamisessa. 2. p. Helsinki: Gaudeamus. 165 s. ISBN 951-662-239-9

Kalamboukis, T. Z. 1995. Suffix stripping with modern Greek. Program, vol. 29, nro 3, s. 313 - 321.

Karetnyk, D., Karlsson, F. & Smart, G. 1991. Knowledge-based indexing of morpho-syntactically analysed language. Expert Systems for Information Management, vol. 4, nro 1, s. 1 - 29.

Karlsson, F. 1982. Johdatusta yleiseen kielitieteeseen. 4. p. Helsinki: Gaudeamus. 279 s. ISBN 951-570-136-8

Karlsson, F. (toim.) 1985. Computational morphosyntax: Report on research 1981 - 84. Helsinki: University of Helsinki, Department of General Linguistics. (Publications, 13.) 178 s. ISBN 951-45-3691-6

Karlsson, F. 1987. Finnish grammar. Porvoo: WSOY. 222 s. ISBN 951-0-11627-0

Karlsson, F. 1990. Constraint grammar as a framework for parsing running text. Teoksessa: Karlgren, H. (toim.) COLING-90: Papers presented to the 13th International Conference on Computational Linguistics, Helsinki, 1990. Vol. 3. Helsinki: Yliopistopaino. S. 168 - 173. ISBN 952-90-2027-9

Karlsson, F. 1994. Yleinen kielitiede. Helsinki: Yliopistopaino. 302 s. ISBN 951-570-136-8

Keen, E. M. 1991. The effect of stemming strength on the effectiveness of output ranking. Teoksessa: Jones, K. P. (toim.) The structuring of information: proceedings of Informatics 11 conference. University of York, 20 - 22 March 1991. London: Aslib. S. 37 - 50. ISBN 0-85142-282-9

Keen, E. M. 1992. Presenting results of experimental retrieval comparisons. Information Processing and Management, vol. 28, nro 4, s. 491 - 502.

Kekäläinen, J. 1999. The effects of query complexity, expansion and structure on retrieval performance in probalistic text retrieval. Tampere: Tampereen yliopisto. (Acta Universitatis Tamperensis 678.) 170 s. ISBN 951-44-4596-1

Kinnucan, M. T., Nelson, M. J. & Allen, B. L. 1987. Statistical methods in information science research. Teoksessa: Williams, M.E. (toim.). Annual Review of Information Science and Technology, vol. 22. Amsterdam: Elsevier. S. 147 - 178. ISBN 0-444-70302-0

Kohonen, T. 1995. Self-organizing maps. Berlin: Springer. 362 s. ISBN 3-540-58600-8

Koskenniemi, K. 1983. Two-level morphology: A general computational model for word-form recognition and production. Helsinki: University of Helsinki, Department of General Linguistics. (Publications, 11.) 160 s. ISBN 951-45-3201-5

Koskenniemi, K. 1985a. An application of the two-level model to Finnish. Teoksessa: Karlsson, F. (toim.). Computational morphosyntax: Report on research 1981 - 84. Helsinki: University of Helsinki, Department of General Linguistics. (Publications, 13.) S. 19 - 41. ISBN 951-45-3691-6

Koskenniemi, K. 1985b. FINSTEMS: A module for information retrieval. Teoksessa: Karlsson, F. (toim.). Computational morphosyntax: Report on research 1981 - 84. Helsinki: University of Helsinki, Department of General Linguistics. (Publications, 13.) S. 81 - 92. ISBN 951-45-3691-6

Koskenniemi, K. 1990. Finite-state parsing and disambiguation. Teoksessa: Karlgren, H. (toim.) COLING-90: Papers presented to the 13th International Conference on Computational Linguistics, Helsinki, 1990. Vol. 2. Helsinki: Yliopistopaino. S. 229 - 232. ISBN 952-90-2026-0

Kotzias, K. 1990. How to respond to different language particularities by indexing texts using automatic text analysis. Proceedings of the 14th International Online Information Meeting, London, 11 - 13 Dec. 1990. Oxford: Learned Information. S. 61 - 68.

Kristensen, J. 1993. Expanding end-users' query statements for free text searching with a search-aid thesaurus. Information Processing and Management, vol. 29, nro 6, s. 733 - 744.

Kristensen, J. 1995. Aiherelevanssi ja käyttäjärelevanssi tulkinnan näkökulmasta. Kirjastotiede ja informatiikka, vol. 14, nro 3, s. 95 - 99.

Kristensen, J. & Järvelin, K. 1990. The effectiveness of a searching thesaurus in free-text searching in a full-text database. International Classification, vol. 17, nro 2, s. 77 - 84.

Krovetz, R. 1993. Viewing morphology as an inference process. Proceedings of the Sixteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. Pittsburg, PA: 27 June - 1 July 1993. New York: The Association for Computing Machinery. S. 191 - 202. ISBN 0-89791-605-0

Laalo, K. 1989. Homonymiasta ja polysemiasta. Virittäjä, vol. 93, vihko 2, s. 220 - 235.

Laalo, K. 1990. Säkeistä patoihin: Suomen kielen monitulkintaiset sanamuodot. Helsinki: Suomalaisen Kirjallisuuden Seura. 113 s. ISBN 951-717-638-4

Lahtinen, T. 1995. Guidelines for extracting index terms. Teoksessa: Koskenniemi, K. (toim.). Abstracts of posters presented at the 10th Nordic conference of computational linguistics, NODALIDA-95. Helsinki, 29 - 30 May 1995. Helsinki: Yliopistopaino. S. 17 - 20.

Lancaster, F. W. 1986. Vocabulary control for information retrieval. 2nd ed. Arlington, VA: Information Resources Press. 270 s. ISBN 0-87815-053-6

Lange, H. R. 1993. Speech synthesis and speech recognition: tomorrow's human-computer interfaces?. Teoksessa: Williams, M.E. (toim.). Annual Review of Information Science and Technology, vol. 28. Medford, NJ: Learned Information. S. 153 - 185. ISBN 0-938734-75-X

Ledwith, R. 1992. On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases. Information Processing and Management, vol. 28, nro 4, s. 451 - 455.

Lehti luukussa? Lehti ja lehtitekniikka muuttuvassa viestintämaisemassa 1989. Helsinki: Liikenneministeriö. 169 s. ISBN 951-861-651-5

Leino, P. 1991. Kieleen mieltä: hyvää suomea. Helsinki: Otava. 477 s. ISBN 951-1-12024-7

Lennon, M., Peirce, D. S., Tarry, B. D. & Willett, P. 1981. An evaluation of some conflation algorithms for information retrieval. Journal of Information Science, vol. 3, s. 177 - 183.

Leppänen, E. 1996. Homografiongelma tekstihaussa ja homografien disambiguoinnin vaikutukset. Informaatiotutkimus, vol. 15, nro 4, s. 133 - 144.

Liddy, E. D. 1990. Anaphora in natural language processing and information retrieval. Information Processing and Management, vol. 26, nro 1, s. 39 - 52.

Liddy, E., Bonzi, S., Katzer, J. & Oddy, E. 1987. A study of discourse anaphora in scientific abstracts. Journal of the American Society for Information Science, vol. 38, nro 4, s. 255 - 261.

Newton, S. J. 1983. Text filing and retrieval systems: A practical evaluation guide. Manchester: NCC Publications. 95 s. + liitt. 34 s. ISBN 0-85012-394-1

Paice, C. D. 1990. Another stemmer. ACM Sigir Forum, vol. 24, nro 3, s. 56 - 61.

Paice, C. D. 1994. An evaluation method for stemming algorithms. Teoksessa: Croft, W. B. & van Rijsbergen, C. J. (toim.) Proceedings of the 17th International Conference on Research and Development in Information Retrieval (ACM-SIGIR). 3 - 6 July 1994, Dublin. London: Springer-Verlag. S. 42 - 50. ISBN 3-540-19889-X

Pennanen, M. & Vakkari, P. 2000. Ongelman jäsentymisen yhteys tiedonhaun muutoksiin tehtäväprosessin aikana. Informaatiotutkimus, vol. 19, nro 1, s. 3 - 10.

Penttilä, A. 1975. Homonüümiast, eriti soome keelt silmas pidades. Congressus Tertius Internationalis Fenno-Ugristarum Tallinnae Habitus. Pars I. Tallinn. S. 322 - 326.

Pirkola, A. 1999. Studies on linguistic problems and methods in text retrieval. Tampere: University of Tampere. 99 s. + liitt. 92 s. (Acta Universitatis Tamperensis 672.) ISBN 951-44-4582-1

Pirkola, A. & Järvelin, K. 1996a. The effect of anaphor and ellipsis resolution on proximity searching in a text database. Information Processing & Management, vol. 32, nro 2, s. 199 - 216.

Pirkola, A. & Järvelin, K. 1996b. Recall and precision effects of anaphor and ellipsis resolution in proximity searching in a text database. Teoksessa: Ingwersen, P. & Pors, N. O. (toim.) Proceedings of CoLIS, 2nd International Conference on Conceptions of Library and Information Science: Integration in Perspective. 13 - 16 Oct. 1996, Copenhagen. Copenhagen: The Royal School of Librarianship. S. 459 - 475. ISBN 87-7415-260-2

Popovic, M. & Willett, P. 1992. The effectiveness of stemming for natural-language access to Slovene textual data. Journal of the American Society for Information Science, vol. 43, nro. 1, s. 384 - 390.

Pors, N. O. 2000. Information retrieval, experimental models and statistical analysis. Journal of Documentation, vol. 56, nro 1, s. 55 - 70.

Porter, M. F. 1980. An algorithm for suffix stripping. Program, vol. 14, nro 3, s. 130 - 137.

Rau, L. F. 1991. Extracting company names from text. Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications. Miami Beach, FL, 24 - 28 Feb. 1991. Los Alamitos: IEEE Computer Soc. Press. S. 29 - 32. ISBN 0-8186-2135-4

Regazzi, J. J. 1988. Performance measures for information retrieval systems - An experimental approach. Journal of the American Society for Information Science, vol. 39, nro 4, s. 235 - 251.

Renouf, A. 1993. Sticking to the text: a corpus linguist's view of language. Aslib Proceedings, vol. 45, nro 5, s. 131 - 136.

Robertson, S. E. & Hancock-Beaulieu, M. M. 1992. On the evaluation of IR systems. Information Processing and Management, vol. 28, nro 4, s. 457 - 466.

Ruge, G., Schwarz, C. & Warner, A. 1991. Effectiveness and efficiency in natural language processing for large amounts of text. Journal of the American Society for Information Science, vol. 42, nro 6, s. 450 - 456.

Saffady, W. 1989. Text storage and retrieval systems: A technology survey and product directory. Westport: Meckler. 131 s. ISBN 0-88736-526-42

Salton, G. 1989. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Reading: Addison-Wesley. 530 s. ISBN 0-201-12227-8

Salton, G. & McGill, M. 1983. Introduction to modern information retrieval. Singapore: McGraw-Hill. 448 s. ISBN 0-07-Y66526-5

Sanderson, M. & Crestani, F. 1998. Mixing and merging for spoken document retrieval. Teoksessa: Nikolaou, C. & Stephanidis, C. (toim.) Research and Advanced Technology for Digital Libraries. Second European conference; proceedings. Heraklion, Greece, 21 - 23 Sept. 1998. Berlin: Springer-Verlag. S. 397 -407. ISBN 3-540-65101-2

Saukkonen, P. 1973. Suomen kielen yhdyssanojen rakenne. Teoksessa: Commentationes Fenno-Ugricae: In honorem Erkki Itkonen sexagenarii. Helsinki: Suomalais-ugrilainen seura. (Toimituksia 150.) S. 332 - 339.

Saukkonen, P. et al. 1979. Suomen kielen taajuussanasto. Porvoo: WSOY. 536 s. ISBN 951-0-09060-3

Savoy, J. 1993. Stemming of French words based on grammatical categories. Journal of the American Society for Information Science, vol. 44, nro. 1, s.1 - 9.

Savoy, J. 1999. A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, vol. 50, nro. 10, s. 944 - 952.

Schamber, L., Eisenberg, M. B. & Nilan, M. S. 1990. A re-examination of relevance: Toward a dynamic, situational definition. Information Processing & Management, vol 26, nro 6, s. 755 - 776.

Schwarz, C. 1990. Automatic syntactic analysis of free text. Journal of the American Society for Information Science, vol. 41, nro 6, s. 408 - 417.

Siegel, S. & Castellan, N. J. Jr. 1988. Nonparametric statistics for the behavioral sciences. Second edition. Singapore: McGraw-Hill. 399 s. ISBN 0-07-100326-6

Snow, B. 1986. What jargon is really necessary when teaching (and learning) online skills? Online, vol. 10, nro 4, s. 100 - 107.

Spink, A., Greisdorf, H. & Bateman, J. 1998. From highly relevant to non relevant: Examining different regions of relevance. Information Processing & Management, vol 34, nro 5, s. 599 - 621.

Sormunen, E. 1989. An analysis of online searching knowledge for intermediary systems. Espoo: Valtion teknillinen tutkimuskeskus. (VTT Tutkimuksia 630.) 81 s. + liitt. 16 s. ISBN 951-38-3502-2

Sormunen, E. 1994. Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanomalehtiaineistoa sisältävässä tekstikannassa. Espoo: Valtion teknillinen tutkimuskeskus. (VTT Julkaisuja 790.) 162 s. + liitt. 60 s. ISBN 951-38-4138-3

Sormunen, E. 2000. A method for measuring wide range performance of Boolean queries in full-text databases. Tampere: Tampereen yliopisto. (Acta Electronica Universitatis Tamperensis 34.) 216 s. + liitt. 14 s. Osoitteessa: <URL: http://acta.uta.fi>. ISBN 951-44-4732-8

Sormunen, E. & Alkula, R. 1990. Suomenkielisten tekstitietokantojen tallennus- ja hakutekniikkojen kehittäminen: Esitutkimusraportti. Espoo:

Valtion teknillinen tutkimuskeskus. (VTT Tiedotteita 1121.) 53 s. ISBN 951-38-3691-6

Sparck Jones, K. 1974. Automatic indexing. Journal of Documentation, vol. 30, nro 4, s. 393 - 432.

Sparck Jones, K. 1995. Reflections on TREC. Information Processing and Management, vol. 31, nro 3, s. 291 - 314.

Sparck Jones, K. 2000. Further reflections on TREC. Information Processing and Management, vol. 36, nro 1, s. 3 - 34.

Sparck Jones, K. & Kay, M. 1973. Linguistics and information science. New York: Academic Press. (FID Publication 492.) 244 s. ISBN 0-12-656250-4

Sparck Jones, K., Jones, G. J. F., Foote, J. T. & Young, S. J. 1996. Experiments in spoken document retrieval. Information Processing and Management, vol. 32, nro 4, s. 399 - 417.

Swanson, D. R. 1977. Information retrieval as a trial-and-error process. Library Quarterly, vol. 47, nro 2, s. 128 - 148.

Tague-Sutcliffe, J. 1992. The pragmatics of information retrieval experimentation, revisited. Information Processing and Management, vol. 28, nro 4, s. 467 - 490.

Tenopir, C. & Ro, J. S. 1990. Full text databases. Westport: Greenwood Press. (New directions in information management, 21.) 251 s. ISBN 0-87287-709-4

Thönssen, B. 1988. Automatische Indexierung und Schnittstellen zu Thesauri. Nachrichten für Dokumentation, vol. 39, nro 4, s. 227 - 230.

Tietotekniikan sanasto 1990. Helsinki: Tietosanoma. 565 s. ISBN 951-885-013-5.

Tolle, K. M. & Chen, H. 2000. Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American Society for Information Science, vol. 51, nro 4, s. 352 - 370.

Ullman, J. D. 1988. Principles of database and knowledge-base systems. Vol. 1. Rockville, Maryland: Computer Science Press. (Principles of Computer Science Press, 14.) 631 s. ISBN 0-88175-188-X.

Ulmschneider, J. E. & Doszkocs, T. 1983. A practical stemming algorithm for online search assistance. Online Review, vol. 7, nro 4, s. 301 - 318.

Vakkari, P. 1999. Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. Information Processing and Management, vol. 35, nro 6, s. 819 - 837.

Vickery, A. 1988. The experience of building expert search systems. Proc. 12th Int. Online Inf. Meet. Volume 1. London: 6 - 8 Dec. 1988. Oxford: Learned Information. S. 301 - 313. ISBN 0-904-933-68-7

Vickery, A. 1989. Intelligent interfaces for online searching. Aslib Information, vol. 17, nro 11/12, s. 271 - 274.

Voorhees, E. & Harman, D. 2000. Overview of the sixth text retrieval conference (TREC-6) Information Processing and Management, vol. 36, nro 1, s. 3 - 35.

Voutilainen, A. 1994. Three studies of grammar-based surface parsing of unrestricted English text. Helsinki: University of Helsinki, Department of General Linguistics. (Publications 24.) 36 s. ISBN 951-45-6670-X

Walker, S. 1988. Improving subject access painlessly: recent work on the Okapi online catalogue projects. Program, vol. 22, nro 1, s. 21 - 31.

Warner, A. 1991. Quantitative and qualitative assessments of the impact of linguistic theory on information science. Journal of the American Society for Information Science, vol 42, nro 1, s. 64 - 71.

Wormell, I. 1984. Cognitive aspects in natural language and free-text searching. Social Science Information Studies, vol. 4, nro 2 & 3, 131 - 141.

Zobel, J., Moffat, A., Wilkinson, R. & Sacks-Davis, R. 1995. Efficient retrieval of partial documents. Information Processing and Management, vol. 31, nro 3, s. 361 - 377.

PAINAMATTOMAT LÄHTEET

Conexor 2000. Conexor Oy:n WWW-sivut osoitteessa: <URL: http://www.conexor.fi>.

Hjorth, T. 1986. Arkistohakupyyntöjen keräys HS-arkistossa keväällä 1986.

Hjorth, T. 1987a. Arkistokokeilun koehaut.

Hjorth, T. 1987b. Suuren suomenkielisen tekstitietokannan etsintämenetelmät, esimerkkinä sanomalehtiarkisto. Helsingin yliopisto, tietojenkäsittelyopin laitos, laudaturtyö. 104 s.

Inktomi 2000. Web surpasses one billion documents. 18.1.2000. Osoitteessa: <URL: http://www.inktomi.com/new/press/billion.html>

Johnson, S. E., Jourlin, P., Sparck Jones, K. & Woodland, P. C. 1999. Spoken document retrieval for TREC-8 at Cambridge University. Proceedings of the eighth Text Retrieval Conference. Gaithersburg, MD, 17 - 19 Nov. 1999. <URL: http://trec.nist.gov/pubs/trec8/t8_proceedings.html>

Keskustalo, H. 1994. Suomenkielisen tekstitietokannan hakemiston rakentamisesta INQUERY/TWOL-ympäristössä. Tampereen yliopisto, informaatiotutkimuksen laitos, sivuaineen tutkielma. 55 s.

Koskenniemi, K. 1985c. Kirje (ongelmasanojen analyysi kaksitasoleksikon avulla).

Kristensen, J. 1989. Tesauruksen rooli vapaatekstihaussa. Tampereen yliopisto, kirjastotieteen ja informatiikan laitos, pro gradu -tutkielma. 82 s.

Kristensen, J. 1992. Vapaasanahakujen laajentaminen hakutesauruksen avulla haettaessa indeksoimattomasta tekstitietokannasta. Tampereen yliopisto, kirjastotieteen ja informatiikan laitos, lisensiaatintutkimus. 132 s.

Niemistö, J. 1988. Suomenkielisten tekstidokumenttien arkistointijärjestelmä. Teknillinen korkeakoulu, tietotekniikan osasto, diplomityö. 80 s. + liitt. 2 s.

Nurminen, R. 1986. Suomen kieltä analysoivien ohjelmien vaikutus dokumenttien tallennukseen ja hakuun suorakäyttöjärjestelmissä. Tampereen yliopisto, kirjastotieteen ja informatiikan laitos, pro gradu -tutkielma. 89 s.

Pirkola, A. 1996. Ellipsien ja anaforien resoluution vaikutus läheisyysoperaatioilla tehtävien tekstitietokantahakujen tuloksiin. Tampereen yliopisto, informaatiotutkimuksen laitos, lisensiaatintutkimus. 118 s. + liitt. 90 s.

Salminen, R. 1988. BASIS: Käyttöopas. 34 s.

Salminen, R. 1992. Puhelinkeskustelu 16.4.1992.

Smeaton, A. F. 1991. Natural language processing and information retrieval. Tutorial of the 14th International Conference on Research and Development in Information Retrieval (ACM-SIGIR). October 1991. Chicago. 83 s.

Strzalkowski, T., Lin, F. & Perez-Carballo, J. 1997. Natural language information retrieval: TREC-6 report. Proceedings of the sixth Text Retrieval Conference. Gaithersburg, MD, 19 - 21 Nov. 1997. Osoitteessa: <URL: http://trec.nist.gov/pubs/trec6/t6_proceedings.html>

Strzalkowski, T., Stein, G., Wise, G. B., Perez-Carballo, J., Tapanainen, P., Järvinen, T., Voutilainen, A. & Karlgren, J. 1998. Natural language information retrieval: TREC-7 report. Proceedings of the seventh Text Re-

trieval Conference. Gaithersburg, MD, 9 - 11 Nov. 1998. Osoitteessa: <URL: http://trec.nist.gov/pubs/trec7/t7_proceedings.html>

Ylinen, M. 1991. Kustannus Oy Aamulehden tekstiarkistosta. Softprod 91. Tampere, 23.-25.10.1991. Tietotekniikan liitto. 5 s.

LIITTEET

1 BASIS-hakujärjestelmän testikysymykset
2 FULLTEXT-projektin kyselyt
3 Ohjeita relevanssiarvioijille
4 Esimerkki relevanssiarvoijille annetusta artikkelista
5 Relevanttien artikkeleitten määrä
6 Esimerkki tulosjoukkojen vertailutaulukosta
7 Tulosjoukkojen koon vaihtelu, JA-operaattori
8 Tulosjoukkojen koon vaihtelu, virkeoperaattori
9 Saannin vaihtelu, JA-operaattori
10 Saannin vaihtelu, virkeoperaattori
11 Tarkkuuden vaihtelu, JA-operaattori
12 Tarkkuuden vaihtelu, virkeoperaattori
13 Esimerkki merkitsevyystestilaskelmasta
14 Merkitsevyystestien tulosten yhteenveto
15 Ongelmakyselyjen osumat

LIITE 1

BASIS-HAKUJÄRJESTELMäN TESTIKYSYMYKSET

Vakiokyselyt

Alkuperäisistä 30 kysymyksestä jätettiin teknisistä tms. syistä pois neljä, jotka näkyvät allaolevassa luettelossa pienemmällä kirjasimella. Karsimisen perusteet on selostettu luvun 9 alussa.

1. Leirikoulu

   Kulmakunnalle on tullut koululuokka leirikouluun. Koska sinulla ei ole omia koulukkaita, päätät tutustua aihepiiriin arkistosta löytyvien juttujen avulla, ennen kuin lähdet haastattelureissulle.

2. Lomaosakebisnes

   Suuri rakennusliike aikoo perustaa paikkakunnalle vapaa-ajan kylän, joka on tarkoitus markkinoida lähinnä lomaosakkeina. Minkähänlainen bisnes nämä lomaosakkeet oikein ovat?

3. Maataloustulo (viime vuosilta)

   Eduskunnan budjettiesityksen yhteydessä on tullut kähinää maataloustulosta. Ajattelit tällä kertaa kirjoittaa lyhyen rutiiniuutisen sijasta vähän laajemman katsauksen aiheeseen ja sen ongelmiin.

4. Arvopaperimarkkinalaki

   Osakemarkkinoilla on taas töppäilty ja kirjoitat asiasta uutista. Suomessahan on melko tuore arvopaperimarkkinalaki - mitähän asioita sen piiriin kuuluu, liittyisiköhän se jotenkin tähän asiaan?

5. Valtionyhtiöiden pörssiin meno

   Valtionyhtiöiden omistuspohjan laajentamisesta keskustellaan taas - mitähän aiemmin tuumittiin niiden menemisestä pörssiin?

6. Pohjavesien tutkimukset

   Naapurikunnan sahan lähettyviltä on löydetty pohjavedestä kloorifenolia. Päätät perehtyä pohjavesien tutkimiseen arkistosta löytyvien juttujen avulla ennen kuin otat yhteyttä vesipiirin tutkijaan.

7. Metsäteollisuuden investoinnit

   Paikkakunnalla vieraileva ministeri sanoo puheessaan metsäteollisuuden voittojen menneen palkkaliukumiin. Mielestäsi alalla on kyllä investointejakin tehty, joten katsot, mitä tietoja

arkistosta löytyy.

8. Kirkkojen rakentaminen

   Paikallinen seurakunta on aloittanut uuden kirkon rakentamisen tulipalossa tuhoutuneen sijalle. Seurakuntatoimittajanne on lomalla ja sinut määrätään kirjoittajaksi. Katsot varmuuden vuoksi mallia arkistosta löytyvistä jutuista.

9. Lapset ja mainonta

   Uutta tuotetta X markkinoidaan massiivisella kampanjalla, joka näyttää tehoavan ainakin sinun lapseesi. Hermostuneena lapsen kärttämisestä päätät purkaa kiukkuasi kirjoittamalla lapsista ja mainonnasta kriittisen jutun.

10. Armahdus-keskustelu

    Presidentti on armahtanut rikollisen ja asiaa on tälläkin kertaa arvosteltu rankasti. Tutkit arkistosta, mitä presidentistä ja hänen armahdusoikeudestaan on aiemmin kirjoitettu.

11. Meluntorjunta

    Valtatien levennys naapurikunnassa on tienvarsiasukkaiden mielestä lisännyt olennaisesti tiestä koituvia meluhaittoja ja he vaativat meluaidan rakentamista. Mietit juttua kirjoittaessasi meluntorjuntaa yleisemminkin ja etsit siitä tietoja arkistosta.

12. Tietosuoja

    Soitat virastoon tarkistaaksesi tiedot eräästä henkilöstä (joka ei suostu haastateltavaksi). Puhelimessa oleva henkilö vetoaa tietosuojalakiin eikä anna sinulle tietoja. Päätät vähän vilkaista arkistosta, mitä tietosuojalla oikeastaan tarkoitetaan.

13. Autoverotus

    Taloustoimitus on päättänyt koota teemasivut verotuksesta. Sinun tehtäväksesi on delegoitu katsaus autoverotuksen koukeroihin...

14. Kuntaliitokset

    Valtio patistelee kotikuntaasi ja sen naapuria liittymään yhteen. Molempien kuntien valtuustoissa kinastellaan kiivaasti asiasta. Mitenkähän asian käsittely on sujunut muissa vastaavissa kunnissa?

15. Seurakuntavaalit

    Seurakuntavaalit ovat taas tulossa. Olet tekemässä juttua niiden valmisteluista ja ehdokkaista. Mitähän teemoja aiemmissa vaaleissa on ollut esillä?

16. Hintavalvontalaki

    Sinä ja kollegasi kinastelette siitä, mitä hintavalvontalain alaan kuuluu. Rauhanomaisena

henkilönä etsit kättä pidempää eli tarkempaa tietoa talon arkistosta.

17. Korkotukilainat

Ekonomistit pohdiskelevat korkotukilainojen kohtaloa. Koska et muista varmasti, mitkä lainat valtion korkotukea saavatkaan, vilkaiset aiempia juttuja aiheesta.

(18. Maan pakkolunastus)

Kunnan teollisuusaluetta pitäisi laajentaa, mutta rajanaapurit eivät halua myydä lisämaata. Kunnanjohtaja esittää pakkolunastusta. Mitähän asioita maan pakkolunastukseen yleensä liittyy?

19. Maatalouden vientituki

Hallituspuolueet ja oppositio kiistelevät, kenen maksettavaksi maatalouden vientituki pannaan. Lehtesi lukijakunta on aiemmin kommentoinut aihepiiriin liittyviä juttujasi kipakasti (ja aamuyöllä), joten nyt aiot esittää asian riittävän monipuolisesti.

20. Wärtsilän tilintarkastus

Maakunnallisen suuryrityksen konkurssin tiimoilta on noussut kysymys firman tilintarkastuksen tasosta. Kohu tuntuu samantapaiselta kuin Wärtsilän konkurssin yhteydessä, joten haluat tutustua tapaus Wärtsilään tarkemmin.

21. Veijo Meri

Veijo Meri on saanut (jälleen) kirjallisuuspalkinnon. Tiedote asiasta on kovin lyhyt, joten etsit arkistostakin lisätietoja Merestä.

(22. Tuntematon sotilas)

Paikallinen harrastajateatteri on kunnianhimoisesti ottanut ohjelmistoonsa Tuntemattoman. Mietit, millaisen arvostelun siitä kirjoittaisit ja etsit malliksi aiempien näytelmä- ja elokuvaversioiden arvosteluja.

(23. Sirola-opisto)

Sirola-opiston kiinteistön huhutaan olevan myynnissä. Millainen opisto oikein onkaan kyseessä?

24. Neste Oy; maakaasu

Teollisuus kärttää taas ratkaisua lisääntyvään energiantarpeeseensa. Yksi vaihtoehdoista olisi maakaasu; selvität, millaisia suunnitelmia Nesteellä on ollut sen suhteen.

25. Turo Oy:n eläkesotkut

Turo Oy:n eläkesäätiön vastuista virinnyt oikeusjuttu on nyt saanut ratkaisunsa ylemmässä oikeusasteessa. Löytyykö arkistosta tietoja raastuvanoikeuden käsittelystä?

26. Japanin pääomamarkkinat

Teet juttua Japanin talouselämästä. Pääomamarkkinoista tarvittaisiin vielä lisätietoja.

27. Yksilöllinen varhaiseläke

    Eläkkeiden ikärajojen muuttamisesta on keskusteltu. Teet juttua erikoiseläkkeistä, joiden yksi tyyppi on yksilöllinen varhaiseläke. Et muista kovin tarkkaan sen ominaisuuksia, joten arkistolle on taas asiaa.

28. Metalliteollisuuden suhdannenäkymät

    Teet arviota laman vaikutuksesta metalliteollisuuteen ja hankit malliksi juttuja arkistosta.

29. Työllisyyslaki

    Kunnat valittavat, etteivät pysty täyttämään työllisyyslain velvoitteita ja vaativat niiden poistamista. Tutkit arkistosta, miten työllisyyslakiin on aiemmin suhtauduttu.

(30. Kauppojen aukioloajat)

    Kaupan järjestöt vaativat kauppojen aukioloaikojen täydellistä vapauttamista. Mietit, mitä eri näkökulmia asiaan liittyy ja etsit arkistosta taustatueksi kauppojen aukioloa käsitteleviä juttuja.

Ongelmakyselyt

31. Inga Sulin
32. Tarton rauha
33. Kansan Uutiset
34. Vuoden kylä
35. Yhtyneet Paperitehtaat
36. Seitsemän veljestä
37. Suomen Pankin korkopolitiikka
38. Hallittu rakennemuutos
39. Muumi
40. Salmén
41. Ihonen
42. Teräväinen
43. Halva
44. Sri Lanka
45. Takinkääntäjät

FULLTEXT-PROJEKTIN KYSELYT

Kysely-tyypin tunnus	Kyselytyypin toteuttamistapa eri tutkimusympäristöissä
	Perinteinen yhdistelmäkysely
ABC	T1: hakijan katkaisemat hakusanat
	Perinteinen osien yhdistelmäkysely
ABCabc	T1: hakijan katkaisemat hakusanat ja hakusanan osat
	- - -
	Hakusana perusmuodossa
A	T4, T5: haetaan hakusana sellaisenaan (peruskysely)
AC	T2, T3, T4: perusmuotoinen hakusana syötetään taivutusvartalo-ohjelmalle, jonka tuottamat vartalot sijoitetaan kyselyyn alkuperäisen hakusanan tilalle; haetaan hakemistoista varsinaiset hakusanat sekä niillä alkavat yhdyssanat (yhdyssanakysely)
	Hakusana ja johdosperhe perusmuodossa
AB	T4, T5: haetaan hakusana ja sen muu johdosperhe sellaisinaan (johdoskysely)
ABC	T2, T3, T4: hakusanan ja johdosperheen perusmuodot syötetään taivutusvartalo-ohjelmalle, sen tuottamilla vartaloilla haetaan näillä alkavat (yhdys)sanat (yhdistelmäkysely)
	Hakusana yhdyssanan eri osina
AC	T5: haetaan hakusanoilla, joihin on merkitty eri yhdyssanan osien tunnukset (yhdyssanakysely)
	Yhdistelmäkysely
ABC	T5: haetaan hakusanoilla ja niiden johdosperheen jäsenillä, joihin on merkitty yhdyssanan osien tunnukset (yhdistelmäkysely)
	Osien perusmuodot
Aa	T4, T5: haetaan hakusana (joka on yhdyssana) ja sen osat sellaisinaan (osien peruskysely)
ACac	T2, T3, T4: hakusanojen perusmuodot ja näiden osat syötetään taivutusvartalo-ohjelmalle, jonka tuottamat vartalot sijoitetaan kyselyyn alkuperäisten hakusanojen tilalle (osien yhdyssanakysely)

Osien perusmuodot ja johdokset

ABab T4, T5: haetaan hakusana (joka on yhdyssana) ja sen osat sekä näiden johdosperheet sellaisinaan (osien johdoskysely)

ABCabc T2, T3, T4: hakusana ja sen osat sekä näiden johdosperheet annetaan syötteeksi vartalo-ohjelmalle, jonka tuottamilla vartaloilla haetaan yhdyssanoja, jotka alkavat hakusanalla, sen osilla tai näiden johdosperheen jäsenellä (osien yhdistelmäkysely)

Osien perusmuodot ja yhdyssanat

ACac T5: haetaan hakusanalla ja sen osilla joihin on merkitty yhdyssanan osien tunnukset (osien yhdyssanakysely)

Osien yhdistelmäkysely

ABCabc T5: haetaan hakusanan osilla ja näiden johdosperheen jäsenillä, joihi on merkitty yhdyssanan osien tunnukset (osien yhdistelmäkysely)

HUOM!
Edellä esitetty jaottelu on sama kuin aiemmassa VTT:n julkaisussa (Alkula & Honkela 1992), vaikka kyselytyyppien tunnukset on merkitty eri tavalla (vain notaatio on muuttunut)

Esimerkki: kysely 13, AUTOVEROTUS

JA = JA-operaattori, JAL = läheisyysoperaattori (virkeoperaattori)

Perinteinen hakutapa:

T1: hakija katkaisee hakusanat itse - yhdistelmäkysely (ABC)
kori1: autovero*
T1: hakija katkaisee hakusanat itse, yhdyssanat jaetaan osiinsa
- osien yhdistelmäkysely (ABCabc)
kori2: auto*
kori3: vero*
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)

Perusmuotohakemiston ja ositetun perusmuotohakemiston yhteiset kyselytyypit
- hakija antaa hakusanat perusmuodossa

T4, T5: Peruskysely (A) - haetaan hakusanan perusmuodolla
kori1: autoverotus
T4, T5: Johdoskysely (AB)
- laajennetaan edellistä kyselyä hakusanojen johdosperheillä
kori2: autoverotus TAI autovero TAI autoverottaminen
T4, T5: Osien peruskysely (Aa) - jaetaan hakusana osiinsa
kori3: auto
kori4: verotus
kori5: (kori1) TAI (kori3 JA kori4)
kori6: (kori1) TAI (kori3 JAL kori4)
T4, T5: Osien johdoskysely (ABab)
- laajennetaan edellistä kyselyä hakusanojen osien johdosperheillä
kori7: verotus TAI vero TAI verottaja TAI verottaminen TAI verottaa
kori8: (kori2) TAI (kori3 JA kori7)
kori9: (kori2) TAI (kori3 JAL kori7)

Vain ositetun perusmuotohakemiston (T5) kyselytyypit
Yhdyssanojen osia haettaessa osiin merkitään katkaisukohdan symboli.

T5: Yhdyssanakysely (AC) - hakusana on yhdyssanan osana
kori10: autoverotus TAI autoverotus- TAI -autoverotus- TAI -autoverotus
T5: Yhdistelmäkysely (ABC)
- laajennetaan edellistä kyselyä hakusanojen johdosperheillä
kori11: (autoverotus TAI autoverotus- TAI -autoverotus- TAI -autoverotus) TAI
(autovero TAI autovero- TAI -autovero- TAI -autovero) TAI
(autoverottaminen TAI autoverottaminen- TAI -autoverottaminen-

TAI -autoverottaminen)
T5: Osien yhdyssanakysely (ACac) - jaetaan hakusana osiinsa ja
haetaan sanat, joissa nämä osat esiintyvät
kori12: auto TAI auto- TAI -auto- TAI -auto
kori13: verotus TAI verotus- TAI -verotus- TAI -verotus
kori14: (kori12) JA (kori13)
kori15: (kori12) JAL (kori13)
T5: Osien yhdistelmäkysely (ABCabc)
- laajennetaan edellistä kyselyä hakusanojen osien johdosperheillä
kori16: (verotus TAI verotus- TAI -verotus- TAI -verotus) TAI
(vero TAI vero- TAI -vero- TAI -vero) TAI
(verottaja TAI verottaja- TAI -verottaja- TAI -verottaja) TAI
(verottaminen TAI verottaminen- TAI -verottaminen- TAI
-verottaminen) TAI
(verottaa TAI verottaa- TAI -verottaa- TAI -verottaa)
kori17: (kori12) JA (kori16)
kori18: (kori12) JAL (kori16)

Perinteisen (T2, T3) ja perusmuotohakemiston (T4) kyselytyypit
Käyttäjä antaa hakusanat perusmuodossa ja ne syötetään taivutusvartaloita tuottavalle ohjelmalle. Korvataan kyselyn alkuperäinen hakusana vartalo-ohjelman tuottamilla vartaloilla.

T2, T3, T4: Yhdyssanakysely (AC) - haetaan hakusanalla alkavat sanat
kori10: autoverotus (-> FINSTEMS)
T2, T3, T4: yhdistelmäkysely (ABC)
- laajennetaan edellistä kyselyä hakusanojen johdosperheillä
kori11: autoverotus (-> FINSTEMS) TAI autovero (-> FINSTEMS) TAI
autoverottaminen (-> FINSTEMS)
T2, T3, T4: Osien yhdyssanakysely (ACac) - jaetaan yhdyssana osiinsa,
haetaan osilla alkavat sanat
kori12: auto (-> FINSTEMS)
kori13: verotus (-> FINSTEMS)
kori14: (kori10) TAI (kori12 JA kori13)
kori15: (kori10) TAI (kori12 JAL kori13)
T2, T3, T4: Osien yhdistelmäkysely (ABCabc)
- laajennetaan edellistä kyselyä hakusanojen osien johdosperheillä
kori16: verotus (-> FINSTEMS) TAI vero (-> FINSTEMS) TAI
verottaja (-> FINSTEMS) TAI
verottaminen (-> FINSTEMS) TAI verottaja (-> FINSTEMS)
kori17: (kori11) TAI (kori12 JA kori16)
kori18: (kori11) TAI (kori12 JAL kori16)

VAKIOKYSELYT

#1: Leirikoulu
Perinteinen yhdistelmäkysely - ABC/T1
kori1: leirikoulu*
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4 - A/T4,T5; AC/T2,T3,T4
kori1: leirikoulu
Hakusana yhdyssanan eri osina - AC/T5
kori2: leirikoulu TAI -leirikoulu TAI -leirikoulu- TAI leirikoulu-

#2: Lomaosakebisnes
Perinteinen yhdistelmäkysely - ABC/T1
kori1: lomaosak*
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: lomaosake
Hakusana yhdyssanan eri osina - AC/T5
kori2: lomaosake TAI -lomaosake TAI -lomaosake- TAI lomaosake-

#3: Maataloustulo (viime vuosilta)
Perinteinen yhdistelmäkysely - ABC/T1
kori1: maataloustulo*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: maatalou*
kori3: tulo*
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: maataloustulo
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori2: maatalous
kori3: tulo
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana yhdyssanan eri osina - AC/T5
kori6: maataloustulo TAI maataloustulo- TAI -maataloustulo- TAI -maataloustulo
Osien perusmuodot ja yhdyssanat - ACac/T5
kori7: maatalous TAI maatalous- TAI -maatalous- TAI -maatalous
kori8: tulo TAI tulo- TAI -tulo- TAI -tulo
kori9: (kori7) JA (kori8)
kori10: (kori7) JAL (kori8)

#4: Arvopaperimarkkinalaki
Perinteinen yhdistelmäkysely - ABC/T1
kori1: arvopaperimarkkinala*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: arvopaperimarkkin*
kori3: laki* TAI lake* TAI lai* TAI laei*
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: arvopaperimarkkinalaki
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori2: arvopaperimarkkina
kori3: laki
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana yhdyssanan eri osina - AC/T5
kori6: arvopaperimarkkinalaki TAI arvopaperimarkkinalaki- TAI -arvopaperimarkkinalaki- TAI -arvopaperimarkkinalaki
Osien perusmuodot ja yhdyssanat - ACac/T5
kori7: arvo TAI arvo- TAI -arvo- TAI -arvo
kori8: paperi TAI paperi- TAI -paperi- TAI -paperi
kori9: markkina TAI markkina- TAI -markkina- TAI -markkina
kori10: laki TAI laki- TAI -laki- TAI -laki
kori11: (kori7) JA (kori8) JA (kori9) JA (kori10)
kori12: (kori7) JAL (kori8) JAL (kori9) JAL (kori10)

#5: Valtionyhtiöiden pörssiin meno
Perinteinen yhdistelmäkysely - ABC/T1
kori1: valtionyhtiö*
kori2: pörss*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori5: valtio*
kori6: yhtiö*
kori7: (kori2) JA (kori5) JA (kori6) -- ja
kori8: (kori2) JA (kori5 JAL kori6) -- ja/jal
kori9: (kori2) JAL (kori5) JAL (kori6) -- jal
kori10: (kori3) TAI (kori7) -- ja
kori11: (kori3) TAI (kori8) -- ja/jal
kori12: (kori4) TAI (kori9) -- jal
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: valtionyhtiö
kori2: pörssi
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori5: valtio
kori6: yhtiö
kori7: (kori2) JA (kori5) JA (kori6) -- ja
kori8: (kori2) JA (kori5 JAL kori6) -- ja/jal
kori9: (kori2) JAL (kori5) JAL (kori6) -- jal
kori10: (kori3) TAI (kori7) -- ja
kori11: (kori3) TAI (kori8) -- ja/jal
kori12: (kori4) TAI (kori9) -- jal
Hakusana yhdyssanan eri osina - AC/T5
kori13: valtionyhtiö TAI valtionyhtiö- TAI -valtionyhtiö- TAI -valtionyhtiö
kori14: pörssi TAI pörssi- TAI -pörssi- TAI -pörssi
kori15: (kori13) JA (kori14)
kori16: (kori13) JAL (kori14)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori17: valtio TAI valtio- TAI -valtio- TAI -valtio
kori18: yhtiö TAI yhtiö- TAI -yhtiö- TAI -yhtiö
kori19: (kori14) JA (kori17) JA (kori18) -- ja
kori20: (kori14) JA (kori17 JAL kori18) -- ja/jal
kori21: (kori14) JAL (kori17) JAL (kori18) -- jal

#6: Pohjavesien tutkimukset
Perinteinen yhdistelmäkysely - ABC/T1
kori1: pohjave*
kori2: tutki*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: pohjavesi
kori2: tutkimus
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T3,T4
kori5: tutkija
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)
Hakusana yhdyssanan eri osina - AC/T5
kori8: pohjavesi TAI pohjavesi- TAI -pohjavesi- TAI -pohjavesi
kori9: tutkimus TAI tutkimus- TAI -tutkimus- TAI -tutkimus
kori10: (kori8) JA (kori9)
kori11: (kori8) JAL (kori9)
Yhdistelmäkysely - ABC/T5
kori12: (tutkimus TAI tutkimus- TAI -tutkimus- TAI -tutkimus) TAI
(tutkiminen TAI -tutkiminen TAI -tutkiminen- TAI tutkiminen-)
TAI (tutkinta TAI -tutkinta TAI -tutkinta- TAI tutkinta-) TAI
TAI (tutkia TAI -tutkia TAI -tutkia- TAI tutkia-) TAI
(tutkiva TAI -tutkiva TAI -tutkiva- TAI tutkiva-) TAI
(tutkija TAI -tutkija TAI -tutkija- TAI tutkija-)
kori13: (kori8) JA (kori12)
kori14: (kori8) JAL (kori12)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori15: pohja TAI pohja- TAI -pohja- TAI -pohja
kori16: vesi TAI vesi- TAI -vesi- TAI -vesi
kori17: (kori9) JA (kori15) JA (kori16) -- ja
kori18: (kori9) JA (kori15 JAL kori16) -- ja/jal
kori19: (kori9) JAL (kori15) JAL (kori16) -- jal
Osien yhdistelmäkysely - ABCabc/T5
kori20: (kori12) JA (kori15) JA (kori16) -- ja
kori21: (kori12) JA (kori15 JAL kori16) -- ja/jal
kori22: (kori12) JAL (kori15) JAL (kori16) -- jal

#7: Metsäteollisuuden investoinnit
Perinteinen yhdistelmäkysely - ABC/T1
kori1: metsäteolli*
kori2: investoi*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: metsäteollisuus
kori2: investointi
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T3,T4
kori5: investointi TAI investoiminen TAI investoida TAI investoiva TAI investoija
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)
Hakusana yhdyssanan eri osina - AC/T5
kori8: metsäteollisuus TAI metsäteollisuus- TAI -metsäteollisuus- TAI -metsäteollisuus
kori9: investointi TAI investointi- TAI -investointi- TAI -investointi
kori10: (kori8) JA (kori9)
kori11: (kori8) JAL (kori9)
Yhdistelmäkysely - ABC/T5
kori12: (investointi TAI investointi- TAI -investointi- TAI -investointi)
TAI (investoiminen TAI -investoiminen
TAI -investoiminen- TAI investoiminen-) TAI
(investoida TAI -investoida TAI -investoida- TAI investoida-) TAI
(investoiva TAI -investoiva TAI -investoiva- TAI investoiva-)
TAI (investoija TAI -investoija TAI -investoija- TAI investoija-)
kori13: (kori8) JA (kori12)
kori14: (kori8) JAL (kori12)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori15: metsä TAI metsä- TAI -metsä- TAI -metsä
kori16: teollisuus TAI teollisuus- TAI -teollisuus- TAI -teollisuus
kori17: (kori9) JA (kori15) JA (kori16) -- ja
kori18: (kori9) JA (kori15 JAL kori16) -- ja/jal
kori19: (kori9) JAL (kori15) JAL (kori16) -- jal
Osien yhdistelmäkysely - ABCabc/T5
kori20: (kori12) JA (kori15) JA (kori16) -- ja
kori21: (kori12) JA (kori15 JAL kori16) -- ja/jal
kori22: (kori12) JAL (kori15) JAL (kori16) -- jal

#8: Kirkkojen rakentaminen
Perinteinen yhdistelmäkysely - ABC/T1
kori1: kirkko* TAI kirko*
kori2: raken*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: kirkko
kori2: rakentaminen
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T3,T4
kori5: rakentaa
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)
Hakusana yhdyssanan eri osina - AC/T5
kori8: kirkko TAI kirkko- TAI -kirkko- TAI -kirkko
kori9: rakentaminen TAI rakentaminen- TAI -rakentaminen- TAI -rakentaminen
kori10: (kori8) JA (kori9)
kori11: (kori8) JAL (kori9)
Yhdistelmäkysely - ABC/T5
kori12: (rakentaminen TAI rakentaminen- TAI -rakentaminen- TAI -rakentaminen)
TAI (rakennus TAI -rakennus TAI -rakennus- TAI rakennus-)
TAI (rakentaja TAI -rakentaja TAI -rakentaja- TAI rakentaja-)
TAI (rakentava TAI -rakentava TAI -rakentava- TAI rakentava-)
TAI (rakentaa TAI -rakentaa TAI -rakentaa- TAI rakentaa-)
kori13: (kori8) JA (kori12)
kori14: (kori8) JAL (kori12)

#9: Lapset ja mainonta
Perinteinen yhdistelmäkysely - ABC/T1
kori1: lapsi* TAI lapse* TAI last*
kori2: maino*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: lapsi
kori2: mainonta
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T3,T4
kori5: mainonta TAI mainostaminen TAI mainostaa TAI mainostus TAI mainos TAI mainostaja TAI mainostava
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)
Hakusana yhdyssanan eri osina - AC/T5
kori8: lapsi TAI lapsi- TAI -lapsi- TAI -lapsi
kori9: mainonta TAI mainonta- TAI -mainonta- TAI -mainonta
kori10: (kori8) JA (kori9)
kori11: (kori8) JAL (kori9)
Yhdistelmäkysely - ABC/T5
kori12: (mainonta TAI mainonta- TAI -mainonta- TAI -mainonta) TAI
(mainostaminen TAI -mainostaminen TAI -mainostaminen- TAI mainostaminen-) TAI
(mainostaa TAI -mainostaa TAI -mainostaa- TAI mainostaa-) TAI
(mainostus TAI -mainostus TAI -mainostus- TAI mainostus-) TAI
(mainos TAI -mainos TAI -mainos- TAI mainos-) TAI
TAI
(mainostava TAI -mainostava TAI -mainostava- TAI mainostava-)
kori13: (kori8) JA (kori12)
kori14: (kori8) JAL (kori12)

#10: Armahdus-keskustelu
Perinteinen yhdistelmäkysely - ABC/T1
kori1: armah*
kori2: president*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: armahdus
kori2: presidentti
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T3,T4
kori5: armahdus TAI armahtaminen TAI armahtaa TAI armahtaja
kori6: (kori2) JA (kori5)
kori7: (kori2) JAL (kori5)
Hakusana yhdyssanan eri osina - AC/T5
kori8: armahdus TAI -armahdus TAI -armahdus- TAI armahdus-
kori9: presidentti TAI presidentti- TAI -presidentti- TAI -presidentti
kori10: (kori8) JA (kori9)
kori11: (kori8) JAL (kori9)
Yhdistelmäkysely - ABC/T5
kori12: (armahdus TAI -armahdus TAI -armahdus- TAI armahdus-) TAI
(armahtaminen TAI -armahtaminen TAI
-armahtaminen- TAI armahtaminen-) TAI
(armahtaa TAI -armahtaa TAI -armahtaa- TAI armahtaa-) TAI
(armahtaja TAI -armahtaja TAI -armahtaja- TAI armahtaja-)
kori13: (kori9) JA (kori12)
kori14: (kori9) JAL (kori12)

#11: Meluntorjunta
Perinteinen yhdistelmäkysely - ABC/T1
kori1: meluntorju*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: melu*
kori3: torju*
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: meluntorjunta
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T3,T4
kori2: meluntorjunta TAI meluntorjuminen TAI meluntorjuja
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4:
kori3: melu
kori4: torjunta
kori5: (kori1) TAI (kori3 JA kori4)
kori6: (kori1) TAI (kori3 JAL kori4)
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori7: torjunta TAI torjuminen TAI torjua TAI torjuja
kori8: (kori2) TAI (kori3 JA kori7)
kori9: (kori2) TAI (kori3 JAL kori7)
Hakusana yhdyssanan eri osina - AC/T5
kori10: meluntorjunta TAI meluntorjunta- TAI -meluntorjunta- TAI -meluntorjunta
Yhdistelmäkysely - ABC/T5
kori11: (meluntorjunta TAI meluntorjunta- TAI -meluntorjunta- TAI -meluntorjunta) TAI (meluntorjuminen TAI meluntorjuminen- TAI -meluntorjuminen- TAI -meluntorjuminen) TAI (meluntorjuja TAI meluntorjuja- TAI -meluntorjuja- TAI -meluntorjuja)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori12: melu TAI melu- TAI -melu- TAI -melu
kori13: torjunta TAI torjunta- TAI -torjunta- TAI -torjunta
kori14: (kori12) JA (kori13)
kori15: (kori12) JAL (kori13)
Osien yhdistelmäkysely - ABCabc/T5
kori16: (torjunta TAI torjunta- TAI -torjunta- TAI -torjunta) TAI (torjuminen TAI torjuminen- TAI -torjuminen- TAI -torjuminen) TAI (torjua TAI torjua- TAI -torjua- TAI -torjua) TAI (torjuja TAI torjuja- TAI -torjuja- TAI -torjuja)
kori17: (kori12) JA (kori16)
kori18: (kori12) JAL (kori16)

#12: Tietosuoja
Perinteinen yhdistelmäkysely - ABC/T1
kori1: tietosuoj*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: tieto* TAI tiedo*
kori3: suoja*
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: tietosuoja
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori2: tieto
kori3: suoja
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori6: suoja TAI suojata TAI suojaus TAI suojaaminen
kori7: (kori1) TAI (kori2 JA kori6)
kori8: (kori1) TAI (kori2 JAL kori6)
Hakusana yhdyssanan eri osina - AC/T5
kori9: tietosuoja TAI tietosuoja- TAI -tietosuoja- TAI -tietosuoja
Osien perusmuodot ja yhdyssanat - ACac/T5
kori10: tieto TAI tieto- TAI -tieto- TAI -tieto
kori11: suoja TAI suoja- TAI -suoja- TAI -suoja
kori12: (kori10) JA (kori11)
kori13: (kori10) JAL (kori11)
Osien yhdistelmäkysely- ABCabc/T5
kori14: (suoja TAI suoja- TAI -suoja- TAI -suoja) TAI
(suojata TAI suojata- TAI -suojata- TAI -suojata) TAI
(suojaus TAI suojaus- TAI -suojaus- TAI -suojaus) TAI
suojaaminen)
kori15: (kori10) JA (kori14)
kori16: (kori10) JAL (kori14)

#13: Autoverotus
- kuvattu liitteen alussa yksityiskohtaisena esimerkkitapauksena

#14: Kuntaliitokset
Perinteinen yhdistelmäkysely - ABC/T1
kori1: kuntaliit*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: kunt* TAI kunn*
kori3: liit*
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: kuntaliitos
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori2: kunta
kori3: liitos
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori6: liitos TAI liittäminen TAI liittää
kori7: (kori1) TAI (kori2 JA kori6)
kori8: (kori1) TAI (kori2 JAL kori6)
Hakusana yhdyssanan eri osina - AC/T5
kori9: kuntaliitos TAI kuntaliitos- TAI -kuntaliitos- TAI -kuntaliitos
Osien perusmuodot ja yhdyssanat - ACac/T5
kori10: kunta TAI kunta- TAI -kunta- TAI -kunta
kori11: liitos TAI liitos- TAI -liitos- TAI -liitos
kori12: (kori10) JA (kori11)
kori13: (kori10) JAL (kori11)
Osien yhdistelmäkysely - ABCabc/T5
kori14: (liitos TAI liitos- TAI -liitos- TAI -liitos) TAI
(liittäminen TAI liittäminen- TAI -liittäminen- TAI -liittäminen)
TAI (liittää TAI liittää- TAI -liittää- TAI -liittää)
kori15: (kori10) JA (kori14)
kori16: (kori10) JAL (kori14)

#15: Seurakuntavaalit
Perinteinen yhdistelmäkysely - ABC/T1
kori1: seurakuntavaal*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: seurakun*
kori3: vaali* TAI vaale*
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: seurakuntavaali
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori2: seurakunta
kori3: vaali
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori6: seurakunta TAI seurakunnallinen
kori7: (kori1) TAI (kori6 JA kori3)
kori8: (kori1) TAI (kori6 JAL kori3)
Hakusana yhdyssanan eri osina - AC/T5
kori9: seurakuntavaali TAI seurakuntavaali- TAI
-seurakuntavaali- TAI -seurakuntavaali
Osien perusmuodot ja yhdyssanat - ACac/T5
kori10: seurakunta TAI seurakunta- TAI -seurakunta- TAI -seurakunta
kori11: vaali TAI vaali- TAI -vaali- TAI -vaali
kori12: (kori10) JA (kori11)
kori13: (kori10) JAL (kori11)
Osien yhdistelmäkysely - ABCabc/T5
kori14: (seurakunta TAI seurakunta- TAI -seurakunta- TAI -seurakunta)
TAI (seurakunnallinen TAI seurakunnallinen- TAI
-seurakunnallinen- TAI -seurakunnallinen)
kori15: (kori11) JA (kori14)
kori16: (kori11) JAL (kori14)

#16: Hintavalvontalaki
Perinteinen yhdistelmäkysely - ABC/T1
kori1: hintavalvontala*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: hintavalvon*
kori3: laki* TAI lake* TAI laei* TAI lai*
kori4: (kori2) JA (kori3)
kori5: (kori2) JAL (kori3)
kori6: hinta* TAI hinto* TAI hinna* TAI hinno*
kori7: valvo*
kori8: (kori3) JA (kori6) JA (kori7)
kori9: (kori3) JAL (kori6) JAL (kori7)
kori10: (kori1) TAI (kori4) TAI (kori8)
kori11: (kori1) TAI (kori5) TAI (kori9)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: hintavalvontalaki
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori2: hintavalvonta
kori3: laki
kori4: (kori2) JA (kori3)
kori5: (kori2) JAL (kori3)
kori6: hinta
kori7: valvonta
kori8: (kori3) JA (kori6) JA (kori7)
kori9: (kori3) JAL (kori6) JAL (kori7)
kori10: (kori1) TAI (kori4) TAI (kori8)
kori11: (kori1) TAI (kori5) TAI (kori9)
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori12: valvonta TAI valvoa TAI valvominen TAI valvoja
kori13: (kori3) JA (kori6) JA (kori12)
kori14: (kori3) JAL (kori6) JAL (kori12)
kori15: (kori1) TAI (kori4) TAI (kori13)
kori16: (kori1) TAI (kori5) TAI (kori14)
Hakusana yhdyssanan eri osina - AC/T5
kori17: hintavalvontalaki TAI hintavalvontalaki- TAI
-hintavalvontalaki- TAI -hintavalvontalaki
Osien perusmuodot ja yhdyssanat - ACac/T5
kori18: hinta TAI hinta- TAI -hinta- TAI -hinta
kori19: valvonta TAI valvonta- TAI -valvonta- TAI -valvonta
kori20: laki TAI laki- TAI -laki- TAI -laki
kori21: (kori18) JA (kori19) JA (kori20)
kori22: (kori18) JAL (kori19) JAL (kori20)

Osien yhdistelmäkysely - ABCabc/T5
kori23: (valvonta TAI valvonta- TAI -valvonta- TAI -valvonta) TAI
(valvoa TAI valvoa- TAI -valvoa- TAI -valvoa) TAI
(valvominen TAI valvominen- TAI -valvominen- TAI -valvominen)
TAI (valvoja TAI valvoja- TAI -valvoja- TAI -valvoja)
kori24: (kori18) JA (kori23) JA (kori20)
kori25: (kori18) JAL (kori23) JAL (kori20)

#17: Korkotukilainat
Perinteinen yhdistelmäkysely - ABC/T1
kori1: korkotukilain*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: korkotu*
kori3: laina* TAI laino*
kori4: (kori2) JA (kori3)
kori5: (kori2) JAL (kori3)
kori6: korko* TAI koro*
kori7: tuki* TAI tuke* TAI tue* TAI tui*
kori8: (kori3) JA (kori6) JA (kori7)
kori9: (kori3) JAL (kori6) JAL (kori7)
kori10: (kori1) TAI (kori4) TAI (kori8)
kori11: (kori1) TAI (kori5) TAI (kori9)

Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: korkotukilaina
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori2: korkotuki
kori3: laina
kori4: (kori2) JA (kori3)
kori5: (kori2) JAL (kori3)
kori6: korko
kori7: tuki
kori8: (kori3) JA (kori6) JA (kori7)
kori9: (kori3) JAL (kori6) JAL (kori7)
kori10: (kori1) TAI (kori4) TAI (kori8)
kori11: (kori1) TAI (kori5) TAI (kori9)
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori12: tuki TAI tukea TAI tukeminen
kori13: (kori3) JA (kori6) JA (kori12)
kori14: (kori3) JAL (kori6) JAL (kori12)
kori15: (kori1) TAI (kori4) TAI (kori13)
kori16: (kori1) TAI (kori5) TAI (kori14)
Hakusana yhdyssanan eri osina - AC/T5
kori17: korkotukilaina TAI korkotukilaina- TAI
-korkotukilaina- TAI -korkotukilaina
Osien perusmuodot ja yhdyssanat - ACac/T5
kori18: korko TAI korko- TAI -korko- TAI -korko
kori19: tuki TAI tuki- TAI -tuki- TAI -tuki
kori20: laina TAI laina- TAI -laina- TAI -laina
kori21: (kori18) JA (kori19) JA (kori20)
kori22: (kori18) JAL (kori19) JAL (kori20)
Osien yhdistelmäkysely - ABCabc/T5
kori23: (tuki TAI tuki- TAI -tuki- TAI -tuki) TAI
(tukea TAI tukea- TAI -tukea- TAI -tukea) TAI
(tukeminen TAI tukeminen- TAI -tukeminen- TAI -tukeminen)
kori24: (kori18) JA (kori23) JA (kori20)
kori25: (kori18) JAL (kori23) JAL (kori20)

#18: Maan pakkolunastus
Perinteinen yhdistelmäkysely - ABC/T1
kori1: maa*
kori2: pakkolunast*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori5: pakko* TAI pako*
kori6: lunast*
kori7: (kori5) JA (kori6)
kori8: (kori5) JAL (kori6)
kori9: (kori1) JA (kori2 TAI kori7) -- ja
kori10: (kori1) JA (kori2 TAI kori8) -- ja/jal
kori11: (kori1) JAL (kori2 TAI kori8) -- jal
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: maa
kori2: pakkolunastus
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T3,T4
kori5: pakkolunastus TAI pakkolunastaminen TAI pakkolunastaa
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori8: pakko
kori9: lunastus
kori10: (kori8) JA (kori9)
kori11: (kori8) JAL (kori9)
kori12: (kori1) JA (kori2 TAI kori10) -- ja
kori13: (kori1) JA (kori2 TAI kori11) -- ja/jal
kori14: (kori1) JAL (kori2 TAI kori11) -- jal
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori15: lunastus TAI lunastaminen TAI lunastaja TAI lunastaa
kori16: (kori8) JA (kori15)
kori17: (kori8) JAL (kori15)
kori18: (kori1) JA (kori5 TAI kori16) -- ja
kori19: (kori1) JA (kori5 TAI kori17) -- ja/jal
kori20: (kori1) JAL (kori5 TAI kori17) -- jal
Hakusana yhdyssanan eri osina - AC/T5
kori21: maa TAI maa- TAI -maa- TAI -maa
kori22: pakkolunastus TAI pakkolunastus- TAI -pakkolunastus- TAI -pakkolunastus
kori23: (kori21) JA (kori22)

kori24: (kori21) JAL (kori22)
Yhdistelmäkysely - ABC/T5
kori25: (pakkolunastus TAI pakkolunastus- TAI -pakkolunastus-
TAI -pakkolunastus) TAI
(pakkolunastaminen TAI pakkolunastaminen- TAI
-pakkolunastaminen- TAI -pakkolunastaminen) TAI
TAI (pakkolunastaa TAI pakkolunastaa- TAI -pakkolunastaa-
TAI -pakkolunastaa)
kori26: (kori21) JA (kori25)
kori27: (kori21) JAL (kori25)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori28: pakko TAI pakko- TAI -pakko- TAI -pakko
kori29: lunastus TAI lunastus- TAI -lunastus- TAI -lunastus
kori30: (kori21) JA (kori28) JA (kori29) -- ja
kori31: (kori21) JA (kori28 JAL kori29) -- ja/jal
kori32: (kori21) JAL (kori28) JAL (kori29) -- jal
Osien yhdistelmäkysely - ABCabc/T5
kori33: (lunastus TAI lunastus- TAI -lunastus- TAI -lunastus) TAI
(lunastaminen TAI lunastaminen- TAI
-lunastaminen- TAI -lunastaminen) TAI
(lunastaja TAI lunastaja- TAI -lunastaja- TAI -lunastaja) TAI
(lunastaa TAI lunastaa- TAI -lunastaa- TAI -lunastaa)
kori34: (kori21) JA (kori28) JA (kori33) -- ja
(kori21) JA (kori28 JAL kori33) -- ja/jal
(kori21) JAL (kori28) JAL (kori33) -- jal

#19: Maatalouden vientituki
Perinteinen yhdistelmäkysely - ABC/T1
kori1: maatalou*
kori2: vientitu*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori5: vienti* TAI vienni*
kori6: tuki* TAI tuke* TAI tue*
kori7: (kori5) JA (kori6)
kori8: (kori5) JAL (kori6)
kori9: (kori1) JA (kori2 TAI kori7) -- ja
kori10: (kori1) JA (kori2 TAI kori8) -- ja/jal
kori11: (kori1) JAL (kori2 TAI kori8) -- jal
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: maatalous
kori2: vientituki
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori5: vienti
kori6: tuki
kori7: (kori5) JA (kori6)
kori8: (kori5) JAL (kori6)
kori9: (kori1) JA (kori2 TAI kori7) -- ja
kori10: (kori1) JA (kori2 TAI kori8) -- ja/jal
kori11: (kori1) JAL (kori2 TAI kori8) -- jal
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori12: tuki TAI tukea TAI tukija TAI tukeminen
kori13: (kori5) JA (kori12)
kori14: (kori5) JAL (kori12)
kori15: (kori1) JA (kori2 TAI kori13) -- ja
kori16: (kori1) JA (kori2 TAI kori14) -- ja/jal
kori17: (kori1) JAL (kori2 TAI kori14) -- jal

Hakusana yhdyssanan eri osina - AC/T5
kori18: maatalous TAI maatalous- TAI -maatalous- TAI -maatalous
kori19: vientituki TAI vientituki- TAI -vientituki- TAI -vientituki
kori20: (kori18) JA (kori19)
kori21: (kori18) JAL (kori19)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori22: vienti TAI vienti- TAI -vienti- TAI -vienti
kori23: tuki TAI tuki- TAI -tuki- TAI -tuki
kori24: (kori18) JA (kori22) JA (kori23) -- ja
kori25: (kori18) JA (kori22 JAL kori23) -- ja/jal
kori26: (kori18) JAL (kori22) JAL (kori23) -- jal
Osien yhdistelmäkysely - ABCabc/T5
kori27: (tuki TAI tuki- TAI -tuki- TAI -tuki) TAI
(tukea TAI tukea- TAI -tukea- TAI -tukea) TAI
(tukija TAI tukija- TAI -tukija- TAI -tukija) TAI
(tukeminen TAI tukeminen- TAI -tukeminen- TAI -tukeminen)
kori28: (kori18) JA (kori22) JA (kori27) -- ja
kori29: (kori18) JA (kori22 JAL kori27) -- ja/jal
kori30: (kori18) JAL (kori22) JAL (kori27) -- jal

#20: Wärtsilän tilintarkastus
Perinteinen yhdistelmäkysely - ABC/T1
kori1: Wärtsilä*
kori2: tilintarkast*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori5: tili* TAI tile*
kori6: tarkast*
kori7: (kori5) JA (kori6)
kori8: (kori5) JAL (kori6)
kori9: (kori1) JA (kori2 TAI kori7) -- ja
kori10: (kori1) JA (kori2 TAI kori8) -- ja/jal
kori11: (kori1) JAL (kori2 TAI kori8) -- jal
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: Wärtsilä
kori2: tilintarkastus
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T3,T4
kori5: tilintarkastus TAI tilintarkastaminen TAI tilintarkastaja
kori6: (kori1) JA (kori5)

kori7: (kori1) JAL (kori5)
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori8: tili
kori9: tarkastus
kori10: (kori8) JA (kori9)
kori11: (kori8) JAL (kori9)
kori12: (kori1) JA (kori2 TAI kori10) -- ja
kori13: (kori1) JA (kori2 TAI kori11) -- ja/jal
kori14: (kori1) JAL (kori2 TAI kori11) -- jal
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori15: tarkastus TAI tarkastaa TAI tarkastaminen TAI tarkastaja
kori16: (kori8) JA (kori15)
kori17: (kori8) JAL (kori15)
kori18: (kori1) JA (kori5 TAI kori16) -- ja
kori19: (kori1) JA (kori5 TAI kori17) -- ja/jal
kori20: (kori1) JAL (kori5 TAI kori17) -- jal
Hakusana yhdyssanan eri osina - AC/T5
kori21: tilintarkastus
kori22: (kori1) JA (kori21)
kori23: (kori1) JAL (kori21)
Yhdistelmäkysely - ABC/T5
kori24: tilintarkastus)
TAI (tilintarkastaminen TAI tilintarkastaminen- TAI -tilintarkastaminen- TAI -tilintarkastaminen) TAI (tilintarkastaja TAI tilintarkastaja- TAI -tilintarkastaja- TAI -tilintarkastaja)
kori25: (kori1) JA (kori24)
kori26: (kori1) JAL (kori24)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori27: tili TAI tili- TAI -tili- TAI -tili
kori28: tarkastus TAI tarkastus- TAI -tarkastus- TAI -tarkastus
kori29: (kori1) JA (kori27) JA (kori28) -- ja
kori30: (kori1) JA (kori27 JAL kori28) -- ja/jal
kori31: (kori1) JAL (kori27) JAL (kori28) -- jal
Osien yhdistelmäkysely- ABCabc/T5
kori32: (tarkastus TAI tarkastus- TAI -tarkastus- TAI -tarkastus) TAI (tarkastaa TAI tarkastaa- TAI -tarkastaa- TAI -tarkastaa) TAI (tarkastaminen TAI tarkastaminen- TAI -tarkastaminen- TAI -tarkastaminen)
TAI (tarkastaja TAI tarkastaja- TAI -tarkastaja- TAI -tarkastaja)
kori33: (kori1) JA (kori27) JA (kori32) -- ja
kori34: (kori1) JA (kori27 JAL kori32) -- ja/jal
kori35: (kori1) JAL (kori27) JAL (kori32) -- jal

#21: Veijo Meri
Perinteinen yhdistelmäkysely - ABC/T1
kori1: veijo*
kori2: meri* TAI mere* tai mert*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: veijo
kori2: meri
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)

#22: Tuntematon sotilas
Perinteinen yhdistelmäkysely - ABC/T1
kori1: tuntemat*
kori2: sotila*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: tuntematon
kori2: sotilas
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)

#23: Sirola-opisto
Perinteinen yhdistelmäkysely - ABC/T1
kori1: sirola*
kori2: opisto*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: sirola
kori2: opisto
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana yhdyssanan eri osina - AC/T5
kori5: opisto TAI -opisto TAI -opisto- TAI opisto-
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)

#24: Neste Oy; maakaasu
Perinteinen yhdistelmäkysely - ABC/T1
kori1: neste*
kori2: maakaasu*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: neste
kori2: maakaasu
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana yhdyssanan eri osina - AC/T5
kori5: maakaasu TAI -maakaasu TAI -maakaasu- TAI maakaasu-
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori8: maa TAI maa- TAI -maa- TAI -maa
kori9: kaasu TAI kaasu- TAI -kaasu- TAI -kaasu
kori10: (kori1) JA (kori8) JA (kori9) -- ja
kori11: (kori1) JA (kori8 JAL kori9) -- ja/jal
kori12: (kori1) JAL (kori8) JAL (kori9) -- jal

#25: Turo Oy:n eläkesotkut
Perinteinen yhdistelmäkysely - ABC/T1
kori1: turo*
kori2: eläke* TAI eläkke*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: turo
kori2: eläke
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana yhdyssanan eri osina - AC/T5
kori5: eläke TAI -eläke TAI -eläke- TAI eläke-
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)

#26: Japanin pääomamarkkinat
Perinteinen yhdistelmäkysely - ABC/T1
kori1: japani*
kori2: pääomamarkkin*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori5: pääom*
kori6: markkin*
kori7: (kori5) JA (kori6)
kori8: (kori5) JAL (kori6)
kori9: (kori1) JA (kori2 TAI kori7) -- ja
kori10: (kori1) JA (kori2 TAI kori8) -- ja/jal
kori11: (kori1) JAL (kori2 TAI kori8) -- jal
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: japani
kori2: pääomamarkkina
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori5: pääoma
kori6: markkina
kori7: (kori5) JA (kori6)
kori8: (kori5) JAL (kori6)
kori9: (kori1) JA (kori2 TAI kori7) -- ja
kori10: (kori1) JA (kori2 TAI kori8) -- ja/jal
kori11: (kori1) JAL (kori2 TAI kori8) -- jal
Hakusana yhdyssanan eri osina - AC/T5
kori12: pääomamarkkina TAI -pääomamarkkina
TAI -pääomamarkkina- TAI pääomamarkkina-
kori13: (kori1) JA (kori12)
kori14: (kori1) JAL (kori12)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori15: pää TAI pää- TAI -pää- TAI -pää
kori16: oma TAI oma- TAI -oma- TAI -oma
kori17: markkina TAI markkina- TAI -markkina- TAI -markkina
kori18: (kori15) JA (kori16) JA (kori17)
kori19: (kori15) JAL (kori16) JAL (kori17)
kori20: (kori1) JA (kori18) -- ja
kori21: (kori1) JA (kori19) -- ja/jal
kori22: (kori1) JAL (kori19) -- jal

#27: Yksilöllinen varhaiseläke
Perinteinen yhdistelmäkysely - ABC/T1
kori1: yksilölli*
kori2: varhaiseläk*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: yksilöllinen
kori2: varhaiseläke
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana yhdyssanan eri osina - AC/T5
kori5: yksilöllinen TAI -yksilöllinen TAI -yksilöllinen- TAI yksilöllinen-
kori6: varhaiseläke TAI -varhaiseläke TAI -varhaiseläke- TAI varhaiseläke
kori7: (kori5) JA (kori6)
kori8: (kori5) JAL (kori6)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori9: varhais TAI varhais- TAI -varhais- TAI -varhais
kori10: eläke TAI eläke- TAI -eläke- TAI -eläke
kori11: (kori5) JA (kori9) JA (kori10) -- ja
kori12: (kori5) JA (kori9 JAL kori10) -- ja/jal
kori13: (kori5) JAL (kori9) JAL (kori10) -- jal

#28: Metalliteollisuuden suhdannenäkymät
Perinteinen yhdistelmäkysely - ABC/T1
kori1: metalliteolli*
kori2: suhdan*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: metalliteollisuus
kori2: suhdanne
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Hakusana yhdyssanan eri osina - AC/T5
kori5: metalliteollisuus TAI metalliteollisuus-
TAI -metalliteollisuus- TAI -metalliteollisuus
kori6: suhdanne TAI suhdanne- TAI -suhdanne- TAI -suhdanne
kori7: (kori5) JA (kori6)
kori8: (kori5) JAL (kori6)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori9: metalli TAI metalli- TAI -metalli- TAI -metalli
kori10: teollisuus TAI teollisuus- TAI -teollisuus- TAI -teollisuus
kori11: (kori6) JA (kori9) JA (kori10) -- ja
kori12: (kori6) JA (kori9 JAL kori10) -- ja/jal
kori13: (kori6) JAL (kori9) JAL (kori10) -- jal

#29: Työllisyyslaki
Perinteinen yhdistelmäkysely - ABC/T1
kori1: työllisyysla*
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori2: työllis*
kori3: laki* TAI lake* TAI laei* TAI lai*
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: työllisyyslaki
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori2: työllisyys
kori3: laki
kori4: (kori1) TAI (kori2 JA kori3)
kori5: (kori1) TAI (kori2 JAL kori3)
Osien perusmuodot ja johdokset - ABab/T4,T5; ABCabc/T2,T3,T4
kori6: työllisyys TAI työllistäminen TAI työllistää TAI työllistäjä
kori7: (kori1) TAI (kori6 JA kori3)
kori8: (kori1) TAI (kori6 JAL kori3)
Hakusana yhdyssanan eri osina - AC/T5
kori9: työllisyyslaki TAI työllisyyslaki- TAI
-työllisyyslaki-TAI-työllisyyslaki
Osien perusmuodot ja yhdyssanat - ACac/T5
kori10: työllisyys TAI työllisyys- TAI -työllisyys- TAI -työllisyys
kori11: laki TAI laki- TAI -laki- TAI -laki
kori12: (kori10) JA (kori11)
kori13: (kori10) JAL (kori11)
Osien yhdistelmäkysely - ABCabc/T5
kori14: (työllisyys TAI työllisyys- TAI -työllisyys- TAI -työllisyys) TAI
(työllistäminen TAI työllistäminen- TAI -työllistäminen- TAI -
työllistäminen)
TAI (työllistää TAI työllistää- TAI -työllistää- TAI -työllistää)
TAI (työllistäjä TAI työllistäjä- TAI -työllistäjä- TAI -työllistäjä)
kori15: (kori14) JA (kori11)
kori16: (kori14) JAL (kori11)

#30: Kauppojen aukioloajat
Perinteinen yhdistelmäkysely - ABC/T1
kori1: kauppa* TAI kauppo* TAI kaupa* TAI kaupo*
kori2: aukioloaik* TAI aukioloaj*
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Perinteinen osien yhdistelmäkysely - ABCabc/T1
kori5: auki*
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)
Hakusana perusmuodossa - A/T4,T5; AC/T2,T3,T4
kori1: kauppa
kori2: aukioloaika
kori3: (kori1) JA (kori2)
kori4: (kori1) JAL (kori2)
Osien perusmuodot - Aa/T4,T5; ACac/T2,T3,T4
kori5: auki
kori6: (kori1) JA (kori5)
kori7: (kori1) JAL (kori5)
Hakusana yhdyssanan eri osina - AC/T5
kori8: kauppa TAI kauppa- TAI -kauppa- TAI -kauppa
kori9: aukioloaika TAI aukioloaika- TAI -aukioloaika- TAI -aukioloaika
kori10: (kori8) JA (kori9)
kori11: (kori8) JAL (kori9)
Osien perusmuodot ja yhdyssanat - ACac/T5
kori12: auki TAI auki- TAI -auki- TAI -auki
kori13: (kori8) JA (kori12)
kori14: (kori8) JAL (kori12)

ONGELMAKYSELYT

#31: Inga Sulin
Perinteinen yhdistelmäkysely - ABC/T1
kori1: inga*
kori2: sulin*
kori3: (kori1) JAL (kori2)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: inga
kori2: sulin
kori3: (kori1) JAL (kori2)

#32: Tarton rauha
Perinteinen yhdistelmäkysely - ABC/T1
kori1: tarto*
kori2: rauha*
kori3: (kori1) JAL (kori2)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: tarton
kori2: rauha
kori3: (kori1) JAL (kori2)
(lisäksi korjauskysely, jossa hakusanat perusmuodossa)

#33: Kansan Uutiset
Perinteinen yhdistelmäkysely - ABC/T1
kori1: kansan
kori2: uutis*
kori3: (kori1) JAL (kori2)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: kansan
kori2: uutiset
kori3: (kori1) JAL (kori2)
(lisäksi korjauskysely, jossa hakusanat perusmuodossa)

#34: Vuoden kylä
Perinteinen yhdistelmäkysely - ABC/T1
kori1: vuoden
kori2: kylä*
kori3: (kori1) JAL (kori2)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: vuoden
kori2: kylä
kori3: (kori1) JAL (kori2)
(lisäksi korjauskysely, jossa hakusanat perusmuodossa)

#35: Yhtyneet Paperitehtaat
Perinteinen yhdistelmäkysely - ABC/T1
kori1: yhtyne*
kori2: paperiteh*
kori3: (kori1) JAL (kori2)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: yhtyneet
kori2: paperitehtaat
kori3: (kori1) JAL (kori2)
(lisäksi korjauskysely, jossa hakusanat perusmuodossa)

#36: Seitsemän veljestä
Perinteinen yhdistelmäkysely - ABC/T1
kori1: seitsemä*
kori2: velje*
kori3: (kori1) JAL (kori2)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: seitsemän
kori2: veljestä
kori3: (kori1) JAL (kori2)
(lisäksi korjauskysely, jossa hakusanat perusmuodossa)

#37a: Suomen Pankin korkopolitiikka
(kaikki hakusanat yhdistetty toisiinsa virkeoperaattorilla)
Perinteinen yhdistelmäkysely - ABC/T1
kori1: suomen
kori2: pank*
kori3: korkopoli*
kori4: (kori1) JAL (kori2) JAL (kori3)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: suomen
kori2: pankki
kori3: korkopolitiikka
kori4: (kori1) JAL (kori2) JAL (kori3)
Hakusana yhdyssanan eri osina - AC/T5:
kori5: korkopolitiikka TAI korkopolitiikka- TAI -korkopolitiikka- TAI - korkopolitiikka
kori6: (kori1) JAL (kori2) JAL (kori5)
(lisäksi korjauskysely, jossa hakusanat perusmuodossa)

#37b: Suomen Pankin korkopolitiikka
(Sanaliitto Suomen pankki yhdistetty virkeoperaattorilla, muut hakusanat JA-operaattorilla)
Perinteinen yhdistelmäkysely - ABC/T1
kori1: suomen
kori2: pank*
kori3: korkopoli*
kori4: (kori1 JAL kori2) JA (kori3)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: suomen
kori2: pankki
kori3: korkopolitiikka
kori4: (kori1 JAL kori2) JA (kori3)
Hakusana yhdyssanan eri osina - AC/T5
kori5: korkopolitiikka TAI korkopolitiikka- TAI -korkopolitiikka- TAI -korkopolitiikka
kori6: (kori1 JAL kori2) JA (kori5)
(lisäksi korjauskysely, jossa hakusanat perusmuodossa)

#38: Hallittu rakennemuutos
Perinteinen yhdistelmäkysely - ABC/T1
kori1: hallit*
kori2: rakennemuuto*
kori3: (kori1) JAL (kori2)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: hallittu
kori2: rakennemuutos
kori3: (kori1) JAL (kori2)
Hakusana yhdyssanan eri osina - AC/T5
kori4: rakennemuutos TAI rakennemuutos- TAI -rakennemuutos- TAI -rakennemuutos
kori5: (kori1) JAL (kori4)
(lisäksi korjauskysely, jossa hakusanat perusmuodossa)

#39: Muumi
Perinteinen yhdistelmäkysely - ABC/T1
kori1: muumi* TAI muume*
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: muumi

#40: Salmen
Perinteinen yhdistelmäkysely - ABC/T1
kori1: salmen*
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: salmen

#41: Ihonen
Perinteinen yhdistelmäkysely - ABC/T1
kori1: ihonen* TAI ihos*
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: ihonen

#42: Teräväinen
Perinteinen yhdistelmäkysely - ABC/T1
kori1: teräväinen* TAI teräväis*
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: teräväinen

#43: Halva
Perinteinen yhdistelmäkysely - ABC/T1
kori1: halva*
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: halva
Hakusana yhdyssanan eri osina - AC/T5
kori2: halva TAI halva- TAI -halva- TAI -halva

#44: Sri Lanka
Perinteinen yhdistelmäkysely - ABC/T1
kori1: sri
kori2: lanka*
kori3: (kori1) JAL (kori2)
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: sri
kori2: lanka
kori3: (kori1) JAL (kori2)

#45a: Takinkääntäjät
Perinteinen yhdistelmäkysely - ABC/T1
kori1: takinkäänt*
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: takinkääntäjä

#45b: Takinkääntäjät
Perinteinen yhdistelmäkysely - ABC/T1
kori1: takinkäänt*
Hakusana sellaisenaan - A/T4,T5; AC/T2
kori1: takinkääntäjä
Hakusana ja johdosperhe perusmuodossa - AB/T4,T5; ABC/T2,T4
kori2: takinkääntäjä tai takinkääntäminen tai takinkääntö

OHJEITA RELEVANSSIARVIOIJILLE

Taustaa:

Tutkimuksen tarkoituksena on selvittää, kuinka hyvin toimituksen elektronisesta (tietokoneella olevasta) juttuarkistosta pystytään löytämään toimittajan tarvitsemia juttuja.

Tilanne:

Olet toimittaja, jonka pitäisi kirjoittaa juttu tietystä aiheesta. Kaipaat aiheesta taustatietoja. Käytössäsi on elektronisessa muodossa oleva arkisto, josta sinulle voidaan hakea ja tulostaa tästä aiheesta aiemmin kirjoitettuja juttuja taustaartikkeleiksi.

Tehtävä:

Ohessa on 30 juttuaihetta, joista jokaisesta on annettu "otsikko" ja lyhyt kuvaus. Jokaisen mukana on myös nippu arkistosta löydettyjä juttuja, jotka on poimittu arkistosta tuon "otsikon" sisältämien avainsanojen perusteella. Nippu sisältää eri tavoin haettuja juttuja sekaisin; toiset vastaavat tarvetta paremmin ja toiset huonommin. Arvioi kukin arkiston juttu siltä kannalta, että jos olisit itse kirjoittamassa "otsikon" kuvaamasta aiheesta juttua, niin olisiko arkistosta löydetty juttu käyttökelpoinen (eli relevantti) tuossa tilanteessa. Huom. sinun ei siis tarvitse kirjoittaa juttua, vaan pelkästään päätellä, olisiko annetuista jutuista hyötyä. Arvioi jokainen juttu erikseen; älä välitä siitä, millaisia juttuja nipussa sitä ennen on ollut.

Arvioinnissa käytetään kolmiportaista asteikkoa. Juttu vastaa tarvetta:

1) HYVIN	· juttu käsittelee haluttua aihetta ja on mielestäsi erittäin käyttökelpoinen, ts. se sisältää tietoa, jota voisit käyttää taustoittamaan juttuasi
2) JONKIN VERRAN	· jutussa on ehkä mainittu haluttu aihe, mutta siinä ei ole riittävästi tietoa tai aiheeseen viitataan vain ohimennen (esim. toimitussihteerin etusivulle tekemä viittaus sisäsivulla olevaan juttuun)
3) EI LAINKAAN	· jutussa ei käsitellä etsimääsi asiaa lainkaan, osasto tai juttutyyppi on täysin väärä tms. - tositilanteessa todennäköisesti heittäisit tällaisen tarjokkaan suoraan paperikoriin

ESIMERKKI RELEVANSSIARVIOIJILLE ANNETUSTA ARTIKKELISTA

KYSYMYS 3/90

JUTTUNUMERO 19196
JULKAISU PVM 900315
OTSIKKO Suora tuki vientimaksuihin
OSASTO tal

Kesannointitavoitteesta vasta puolet saatu kokoon Hemilä vetosi viljelijöihin alan lisäämiseksi Viljelijät ovat jättäneet tänä vuonna kesannointihakemuksia odotettua vähemmän.

Maatilahallituksen mukaan 200 000 hehtaarin kesannointitavoitteesta on koossa tähän mennessä vasta puolet, ja kesannointisopimusten hakuaika päättyy jo tämän viikon perjantaina.

Maatilahallituksen pääjohtaja Kalevi Hemilä vetosi keskiviikkona viljelijöihin kesannointialan lisäämiseksi. Hän muistutti, että jos tavoitteita ei saavuteta, on edessä maatalouden kannalta paljon vaikeampia tuotannon rajoituksia.

Jos kesannointitavoitteesta toteutuu vain puolet, jää ylimääräistä peltoalaa viljelyyn 100 000 hehtaaria suunniteltua enemmän. Hemilä arvioi, että tästä aiheutuu noin 300 miljoonan kilon ylimääräisen viljan vientitarve, ja tämän ylijäämän vaatimat vientikustannukset maatalous maksaa itse.

Hemilä vertasi vientikustannuksia helmikuun maataloustulosopimuksessa sovittuun 510 miljoonaan suoraan tulotukeen. Suora tuki vientimaksuihin

- Jos kesannointisopimuksia ei saada lisää, menee paljon puhuttu suora tulotuki kokonaan turhiin vientitukiaisiin, hän arvioi.

Jos vapaaehtoiset toimet eivät tuota tulosta, joutuvat viljelijät sopeutumaan pakkokeinoihin, kuten tiukkaan hintapolitiikkaan, velvoitekesannointiin tai markkinoimismaksujen ja muiden maksujen korottamiseen.

Hemilän mielestä viime kesän hyvä sato ja lauha talvi ovat houkutelleet koko peltoalan tehokkaaseen käyttöön. Hänen mukaansa viljelijät ovat kuitenkin unohtaneet, että kesannointi on maatalouden oman edun mukaista. Kesannoinnista saatava palkkio on suoraa tulotukea, jolla korvataan maataloudelle tuotannon supistamisesta aiheutuvat tulonmenetykset.

RELEVANTTIEN ARTIKKELEITTEN MÄÄRÄ

Arvioitujen artikkelien määrä ja relevanteiksi arvioitujen artikkelien määrä hakupyynnöittäin

HAKUPYYNTÖ #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Arvioituja artikkeleita yhteensä	15	14	113	22	22	77	48	154	54	39	4	100	139	181	28
Ei lainkaan hyödylliset	10	5	66	5	18	60	24	147	48	21	3	94	103	151	19
Jonkin verran hyödylliset	5	6	20	11	1	7	5	3	3	11	0	2	29	15	6
Hyvin hyödylliset	0	3	27	6	3	10	19	4	3	7	1	4	7	15	3
Hyödyllisiä artikkeleita yhteensä	**5**	**9**	**47**	**17**	**4**	**17**	**24**	**7**	**6**	**18**	**1**	**6**	**36**	**30**	**9**

HAKUPYYNTÖ #	**16**	**17**	18	**19**	**20**	**21**	22	23	**24**	**25**	**26**	**27**	**28**	**29**	30
Arvioituja artikkeleita yhteensä	49	87	18	34	23	11	15	3	17	2	8	12	16	84	99
Ei lainkaan hyödylliset	42	62	15	15	4	5	15	3	7	0	5	3	5	75	93
Jonkin verran hyödylliset	7	11	2	10	7	6	0	0	5	0	2	3	3	4	1
Hyvin hyödylliset	0	14	1	9	12	0	0	0	5	2	1	6	8	5	5
Hyödyllisiä artikkeleita yhteensä	**7**	**25**	**3**	**19**	**19**	**6**	**0**	**0**	**10**	**2**	**3**	**9**	**11**	**9**	**6**

Hakupyynnöt 18, 22, 23 ja 30 jätettiin pois lopullisesta tutkimusjoukosta

ESIMERKKI TULOSJOUKKOJEN VERTAILUTAULUKOSTA

Hakupyyntö 4: Arvopaperimarkkinalaki
Kaikki tulosjoukot yhdistäen saatiin yhteensä 22 dokumenttia:
Hyvin hyödyllisiä 6 (alla lihavoituina),
jonkin verran hyödyllisiä 11, ei lainkaan 5 (alla suluissa).

Juttu-numero	**T1/ ABC**	**T1/ ABC-abc/ JA**	**T1/ ABC-abc/ virke**	**T2/AC**	**T2/ ACac/ JA**	**T2/ ACac/ virke**	**T3/AC**	**T3/ ACac/ JA**	**T3/ ACac/ virke**
(20615)	1	1	1	1	1	1	1	1	1
19739	1	1	1	1	1	1	1	1	1
19493	1	1	1	1	1	1	1	1	1
18801	1	1	1	1	1	1	1	1	1
18638									
17626									
17526	1	1	1	1	1	1	1	1	1
(12936)									
(12080)									
(11748)		1			1				
10005	1	1	1	1	1	1	1	1	1
9714		1			1				
8865	1	1	1	1	1	1	1	1	1
6660		1			1				
6492	1	1	1	1	1	1	1	1	1
5923	1	1	1	1	1	1	1	1	1
5061	1	1	1	1	1	1	1	1	1
4591	1	1	1	1	1	1	1	1	1
4189		1	1		1	1		1	1
4129	1	1	1	1	1	1	1	1	1
(3914)		1			1				
1295	1	1	1	1	1	1	1	1	1
Osumia:	13	18	14	13	18	14	13	14	14
Saanti:	12/17	15/17	13/17	12/17	15/17	13/17	12/17	13/17	13/17
Tarkkuus:	12/13	15/18	13/14	12/13	15/18	13/14	12/13	13/14	13/14

Juttu-numero	**T4,T5/ A**	**T4/AC**	**T5/AC**	**T4,T5/ Aa/JA**	**T4,T5/ Aa/ virke**	**T4/ ACac/ JA**	**T4/ ACac/ virke**	**T5/ ACac/ JA**	**T5/ ACac/ virke**
(20615)	1	1	1	1	1	1	1	1	1
19739	1	1	1	1	1	1	1	1	1
19493	1	1	1	1	1	1	1	1	1
18801	1	1	1	1	1	1	1	1	1
18638								1	
17626								1	
17526	1	1	1	1	1	1	1	1	1
(12936)								1	
(12080)								1	
(11748)				1		1		1	
10005	1	1	1	1	1	1	1	1	1
9714						1		1	
8865	1	1	1	1	1	1	1	1	1
6660						1		1	
6492	1	1	1	1	1	1	1	1	1
5923	1	1	1	1	1	1	1	1	1
5061	1	1	1	1	1	1	1	1	1
4591	1	1	1	1	1	1	1	1	1
4189						1	1	1	1
4129	1	1	1	1	1	1	1	1	1
(3914)				1		1		1	
1295	1	1	1	1	1	1	1	1	1
Osumia:	13	13	13	15	13	18	14	22	14

Saanti:	12/17	12/17	12/17	12/17	12/17	15/17	13/17	17/17	13/17
Tarkkuus:	12/13	12/13	12/13	12/15	12/13	15/18	13/14	17/22	13/14

TULOSJOUKKOJEN KOON VAIHTELU, JA-operaattori

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

Tutkimusympäristö				**Y**	**Y**	**J**	**J**
ja kyselytyyppi	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Perinteinen (T1), ABC	**8**	**9**	**39**	**13**	**3**	**26**	**33**
Finstems (T2), AC	**8**	**9**	**39**	**13**	**3**	**15**	**29**
Finstems (T2), ABC	8	9	39	13	3	**26**	**33**
Hahmotin (T2), AC	**8**	**9**	**39**	**13**	**3**	**15**	**29**
Hahmotin (T2), ABC	8	9	39	13	3	**26**	**33**
Seulonta (T3), AC	**8**	**9**	**39**	**13**	**3**	**14**	**16**
Seulonta (T3), ABC	8	9	39	13	3	**20**	**19**
Perusmuotohakemisto (T4), A	**7**	**5**	**8**	**13**	**2**	**11**	**25**
Perusmuotohakemisto (T4), AB	7	5	8	13	2	**17**	**29**
Perusmuotohakemisto (T4), AC	**8**	**9**	**39**	**13**	**3**	**15**	**29**
Perusmuotohakemisto (T4), ABC	8	9	39	13	3	**26**	**33**
Ositettu perusmuotoh. (T5), A	**7**	**5**	**8**	**13**	**2**	**11**	**25**
Ositettu perusmuotoh. (T5), AB	7	5	8	13	2	**17**	**29**
Ositettu perusmuotoh. (T5), AC	**8**	**9**	**39**	**13**	**3**	**16**	**30**
Ositettu perusmuotoh. (T5), ABC	8	9	39	13	3	**29**	**34**
Perinteinen (T1), ABCabc			**154**	**18**	**16**		
Finstems (T2), ACac			**154**	**18**	**16**		
Finstems (T2), ABCabc			154	18	16		
Hahmotin (T2), ACac			**154**	**18**	**16**		
Hahmotin (T2), ABCabc			154	18	16		
Seulonta (T3), ACac			**108**	**14**	**15**		
Seulonta (T3), ABCabc			108	14	15		
Perusmuotohak. (T4), Aa			**39**	**15**	**9**		
Perusmuotohak. (T4), ABab			39	15	9		
Perusmuotohak. (T4), ACac			**143**	**18**	**16**		
Perusmuotohak. (T4), ABCabc			143	18	16		
Ositettu perusm.h. (T5), Aa			**39**	**15**	**9**		
Ositettu perusm.h. (T5), ABab			39	15	9		
Ositettu perusm.h. (T5), ACac			**125**	**22**	**21**	**46**	**43**
Ositettu perusm.h. (T5), ABCabc			125	22	21	**77**	**48**

Tutkimusympäristö	**J**	**J**	**J**	**Y+J**		**J**	
ja kyselytyyppi	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Perinteinen (T1), ABC	**145**	**46**	**39**	**2**	**8**	**30**	**21**
Finstems (T2), AC	**49**	**6**	**30**	**2**	**8**	**5**	**20**
Finstems (T2), ABC	**145**	**46**	**39**	**2**	8	**30**	20
Hahmotin (T2), AC	**49**	**6**	**30**	**2**	**8**	**5**	**20**
Hahmotin (T2), ABC	**126**	**46**	**38**	**2**	8	**30**	20
Seulonta (T3), AC	**27**	**6**	**30**	**2**	**8**	**5**	**18**
Seulonta (T3), ABC	**91**	**39**	**38**	**2**	8	**30**	18
Perusmuotohakemisto (T4), A	**27**	**5**	**21**	**1**	**5**	**5**	**20**
Perusmuotohakemisto (T4), AB	**76**	**25**	**33**	**1**	5	**28**	20
Perusmuotohakemisto (T4), AC	**49**	**6**	**30**	**2**	**8**	**5**	**20**
Perusmuotohakemisto (T4), ABC	**145**	**46**	**39**	**2**	8	**30**	20
Ositettu perusmuotoh. (T5), A	**27**	**5**	**21**	**1**	**5**	**5**	**20**
Ositettu perusmuotoh. (T5), AB	**76**	**25**	**33**	**1**	5	**28**	20
Ositettu perusmuotoh. (T5), AC	**49**	**9**	**30**	**2**	**8**	**5**	**20**
Ositettu perusmuotoh. (T5), ABC	**130**	**49**	**38**	**2**	8	**30**	20
Perinteinen (T1), ABCabc				**4**	**111**	**178**	**727**
Finstems (T2), ACac				**2**	**116**	**43**	**38**
Finstems (T2), ABCabc				**4**	**116**	**178**	**105**
Hahmotin (T2), ACac				**2**	**116**	**43**	**38**
Hahmotin (T2), ABCabc				**4**	**116**	**178**	**104**
Seulonta (T3), ACac				**2**	**45**	**32**	**24**
Seulonta (T3), ABCabc				**4**	**72**	**113**	**43**
Perusmuotohak. (T4), Aa				**1**	**38**	**24**	**24**
Perusmuotohak. (T4), ABab				**3**	**50**	**79**	**52**
Perusmuotohak. (T4), ACac				**2**	**116**	**43**	**31**
Perusmuotohak. (T4), ABCabc				**4**	**116**	**178**	**100**
Ositettu perusm.h. (T5), Aa				**1**	**38**	**24**	**24**
Ositettu perusm.h. (T5), ABab				**3**	**50**	**79**	**52**
Ositettu perusm.h. (T5), ACac				**2**	**89**	**39**	**69**
Ositettu perusm.h. (T5), ABCabc				**4**	**108**	**160**	**136**

Tutkimusympäristö		**Y**	**Y**	**Y+J**		
ja kyselytyyppi		**17**	**19**	**20**	**21**	**24**
Perinteinen (T1), ABC		**19**	**19**	**9**	**11**	**11**
Finstems (T2), AC		**19**	**14**	**5**	**11**	**11**
Finstems (T2), ABC		19	14	**9**	11	11
Hahmotin (T2), AC		**19**	**14**	**5**	**11**	**11**
Hahmotin (T2), ABC		19	14	**9**	11	11
Seulonta (T3), AC		**19**	**12**	**5**	**10**	**11**
Seulonta (T3), ABC		19	12	**9**	10	11
Perusmuotohakemisto (T4), A		**19**	**10**	**3**	**8**	**10**
Perusmuotohakemisto (T4), AB		19	10	**8**	8	10
Ositettu perusmuotoh. (T5), A		**19**	**10**	**3**	**8**	**10**
Ositettu perusmuotoh. (T5), AB		19	10	**8**	8	10
Ositettu perusmuotoh. (T5), AC		**19**	**12**	**5**	8	**10**
Ositettu perusmuotoh. (T5), ABC		19	12	**20**	8	10

Tutkimusympäristö		**Y**			
ja kyselytyyppi	**25**	**26**	**27**	**28**	**29**
Perinteinen (T1), ABC	**2**	**2**	**11**	**5**	**5**
Finstems (T2), AC	**2**	**2**	**11**	**5**	**5**
Finstems (T2), ABC	2	2	11	5	5
Hahmotin (T2), AC	**2**	**2**	**11**	**5**	**5**
Hahmotin (T2), ABC	2	2	11	5	5
Seulonta (T3), AC	**2**	**2**	**10**	**1**	**5**
Seulonta (T3), ABC	2	2	10	1	5
Perusmuotohakemisto (T4), A	**2**	**2**	**8**	**4**	**5**
Perusmuotohakemisto (T4), AB	2	2	8	4	5
Perusmuotohakemisto (T4), AC	**2**	**2**	**11**	**5**	**5**
Perusmuotohakemisto (T4), ABC	2	2	11	5	5
Ositettu perusmuotoh. (T5), A	**2**	**2**	**8**	**4**	**5**
Ositettu perusmuotoh. (T5), AB	2	2	8	4	5
Ositettu perusmuotoh. (T5), AC	**2**	**2**	**11**	**9**	**5**
Ositettu perusmuotoh. (T5), ABC	2	2	11	9	5
Perinteinen (T1), ABCabc		**8**			**90**
Finstems (T2), ACac		**8**			**34**
Finstems (T2), ABCabc		8			**90**
Hahmotin (T2), ACac		**8**			**34**
Hahmotin (T2), ABCabc		8			**89**
Seulonta (T3), ACac		**5**			**14**
Seulonta (T3), ABCabc		5			**29**
Perusmuotohak. (T4), Aa		**3**			**13**
Perusmuotohak. (T4), ABab		3			**28**
Perusmuotohak. (T4), ACac		**7**			**34**
Perusmuotohak. (T4), ABCabc		7			**89**
Ositettu perusm.h. (T5), Aa		**3**			**13**
Ositettu perusm.h. (T5), ABab		3			**28**
Ositettu perusm.h. (T5), ACac		**5**	**12**	**16**	**20**
Ositettu perusm.h. (T5), ABCabc		5	12	16	**38**

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

Tutkimusympäristö ja kyselytyyppi	N_p	Keski-arvo	Y N_y	Y Keski-arvo	J N_j	J Keski-arvo
Perinteinen (T1), ABC	26	**20,0**	9	**8,0**	8	**41,3**
Finstems (T2), AC	26	12,5	9	7,0	8	17,6
Finstems (T2), ABC	26	**19,8**	9	**7,4**	8	**41,3**
Hahmotin (T2), AC	26	12,5	9	7,0	8	17,6
Hahmotin (T2), ABC	26	**19,0**	9	**7,4**	8	**38,8**
Seulonta (T3), AC	26	10,8	9	6,8	8	13,1
Seulonta (T3), ABC	26	**16,3**	9	**7,2**	8	**31,0**
Perusmuotohakemisto (T4), A	26	8,9	9	6,1	8	12,3
Perusmuotohakemisto (T4), AB	26	13,5	9	6,7	8	27,1
Perusmuotohakemisto (T4), AC	26	12,4	9	7,0	8	17,6
Perusmuotohakemisto (T4), ABC	26	**19,7**	9	**7,4**	8	**41,3**
Ositettu perusmuotoh. (T5), A	26	8,9	9	6,1	8	12,3
Ositettu perusmuotoh. (T5), AB	26	13,5	9	6,7	8	27,1
Ositettu perusmuotoh. (T5), AC	26	12,7	9	6,8	8	18,3
Ositettu perusmuotoh. (T5), ABC	26	**19,8**	9	**8,4**	8	**41,5**
Perinteinen (T1), ABCabc	14	**106,6**	9	**25,8**		
Finstems (T2), ACac	14	41,6	9	21,9		
Finstems (T2), ABCabc	14	**62,0**	9	**25,0**		
Hahmotin (T2), ACac	14	41,6	9	22,0		
Hahmotin (T2), ABCabc	14	**61,9**	9	**25,0**		
Seulonta (T3), ACac	14	25,7	9	15,2		
Seulonta (T3), ABCabc	14	**37,8**	9	**18,2**		
Perusmuotohak. (T4), Aa	14	17,5	9	11,9		
Perusmuotohak. (T4), ABab	14	26,9	9	14,3		
Perusmuotohak. (T4), ACac	14	40,2	9	21,8		
Perusmuotohak. (T4), ABCabc	14	**60,7**	9	**24,9**		
Ositettu perusm.h. (T5), Aa	14	17,5	9	11,9		
Ositettu perusm.h. (T5), ABab	14	26,9	9	14,3		
Ositettu perusm.h. (T5), ACac	14	37,7	9	20,7		
Ositettu perusm.h. (T5), ABCabc	14	**56,6**	9	**25,0**		

TULOSJOUKKOJEN KOON VAIHTELU, virkeoperaattori

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

Tutkimusympäristö				**Y**	**Y**	**J**	**J**
ja kyselytyyppi	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Perinteinen (T1), ABC	**8**	**9**	**39**	**13**	**2**	**7**	**12**
Finstems (T2), AC	**8**	**9**	**39**	**13**	**2**	**3**	**10**
Finstems (T2), ABC	8	9	39	13	2	**7**	**12**
Hahmotin (T2), AC	**8**	**9**	**39**	**13**	**2**	**3**	**10**
Hahmotin (T2), ABC	8	9	39	13	2	**7**	**12**
Seulonta (T3), AC	**8**	**9**	**39**	**13**	**2**	**3**	**1**
Seulonta (T3), ABC	8	9	39	13	2	**5**	**4**
Perusmuotohakemisto (T4), A	**7**	**5**	**8**	**13**	**2**	**3**	**8**
Perusmuotohakemisto (T4), AB	7	5	8	13	2	**6**	**10**
Perusmuotohakemisto (T4), AC	**8**	**9**	**39**	**13**	**2**	**3**	**10**
Perusmuotohakemisto (T4), ABC	8	9	39	13	2	**7**	**12**
Ositettu perusmuotoh. (T5), A	**7**	**5**	**8**	**13**	**2**	**3**	**8**
Ositettu perusmuotoh. (T5), AB	7	5	8	13	2	**6**	**10**
Ositettu perusmuotoh. (T5), AC	**8**	**9**	**39**	**13**	**2**	**3**	**10**
Ositettu perusmuotoh. (T5), ABC	8	9	39	13	2	**7**	**12**
Perinteinen (T1), ABCabc			**56**	**14**	**3**		
Finstems (T2), ACac			**56**	**14**	**3**		
Finstems (T2), ABCabc			56	14	3		
Hahmotin (T2), ACac			**56**	**14**	**3**		
Hahmotin (T2), ABCabc			56	14	3		
Seulonta (T3), ACac			**48**	**14**	**3**		
Seulonta (T3), ABCabc			48	14	3		
Perusmuotohak. (T4), Aa			**10**	**13**	**3**		
Perusmuotohak. (T4), ABab			10	13	3		
Perusmuotohak. (T4), ACac			**34**	**14**	**3**		
Perusmuotohak. (T4), ABCabc			34	14	3		
Ositettu perusm.h. (T5), Aa			**10**	**13**	**3**		
Ositettu perusm.h. (T5), ABab			10	13	3		
Ositettu perusm.h. (T5), ACac			**56**	**14**	**4**	**3**	**12**
Ositettu perusm.h. (T5), ABCabc			56	14	4	**8**	**14**

Tutkimusympäristö	**J**	**J**	**J**	**Y+J**		**J**	
ja kyselytyyppi	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Perinteinen (T1), ABC	**40**	**3**	**28**	**2**	**8**	**30**	**21**
Finstems (T2), AC	**6**	**2**	**20**	**2**	**8**	**5**	**20**
Finstems (T2), ABC	**40**	**3**	**28**	**2**	8	**30**	20
Hahmotin (T2), AC	**6**	**2**	**20**	**2**	**8**	**5**	**20**
Hahmotin (T2), ABC	**35**	**3**	**27**	**2**	8	**30**	20
Seulonta (T3), AC	**3**	**2**	**20**	**2**	**8**	**5**	**18**
Seulonta (T3), ABC	**22**	**3**	**27**	**2**	8	**30**	18
Perusmuotohakemisto (T4), A	**2**	**2**	**12**	**1**	**5**	**5**	**20**
Perusmuotohakemisto (T4), AB	**15**	**3**	**22**	**1**	5	**28**	20
Perusmuotohakemisto (T4), AC	**6**	**2**	**20**	**2**	**8**	**5**	**20**
Perusmuotohakemisto (T4), ABC	**40**	**3**	**28**	**2**	8	**30**	20
Ositettu perusmuotoh. (T5), A	**2**	**2**	**12**	**1**	**5**	**5**	**20**
Ositettu perusmuotoh. (T5), AB	**15**	**3**	**22**	**1**	5	**28**	20
Ositettu perusmuotoh. (T5), AC	**7**	**2**	**20**	**2**	**8**	**5**	**20**
Ositettu perusmuotoh. (T5), ABC	**36**	**3**	**27**	**2**	8	**30**	20
Perinteinen (T1), ABCabc				**2**	**16**	**55**	**138**
Finstems (T2), ACac				**2**	**16**	**12**	**23**
Finstems (T2), ABCabc				**2**	**16**	**55**	**35**
Hahmotin (T2), ACac				**2**	**16**	**12**	**23**
Hahmotin (T2), ABCabc				**2**	**16**	**55**	**34**
Seulonta (T3), ACac				**2**	**10**	**9**	**19**
Seulonta (T3), ABCabc				**2**	**13**	**48**	**22**
Perusmuotohak. (T4), Aa				**1**	**5**	**5**	**20**
Perusmuotohak. (T4), ABab				**1**	**8**	**41**	**25**
Perusmuotohak. (T4), ACac				**2**	**16**	**12**	**22**
Perusmuotohak. (T4), ABCabc				**2**	**16**	**55**	**35**
Ositettu perusm.h. (T5), Aa				**1**	**5**	**5**	**20**
Ositettu perusm.h. (T5), ABab				**1**	**8**	**41**	**25**
Ositettu perusm.h. (T5), ACac				**2**	**12**	**13**	**35**
Ositettu perusm.h. (T5), ABCabc				**2**	**16**	**68**	**48**

Tutkimusympäristö	**Y**	**Y**	**Y**	**Y**	**Y+J**		
ja kyselytyyppi	**15**	**16**	**17**	**19**	**20**	**21**	**24**
Perinteinen (T1), ABC	**3**	**2**	**19**	**7**	**5**	**8**	**6**
Finstems (T2), AC	**3**	**2**	**19**	**7**	**2**	**8**	**6**
Finstems (T2), ABC	3	2	19	7	**5**	8	6
Hahmotin (T2), AC	**3**	**2**	**19**	**7**	**2**	**8**	**6**
Hahmotin (T2), ABC	3	2	19	7	**5**	8	6
Seulonta (T3), AC	**3**	**2**	**19**	**6**	**2**	**8**	**6**
Seulonta (T3), ABC	3	2	19	6	**5**	8	6
Perusmuotohakemisto (T4), A	**3**	**2**	**19**	**6**	**2**	**8**	**3**
Perusmuotohakemisto (T4), AB	3	2	19	6	**5**	8	3
Perusmuotohakemisto (T4), AC	**3**	**2**	**19**	**7**	**2**	8	**6**
Perusmuotohakemisto (T4), ABC	3	2	19	7	**5**	8	6
Ositettu perusmuotoh. (T5), A	**3**	**2**	**19**	**6**	**2**	**8**	**3**
Ositettu perusmuotoh. (T5), AB	3	2	19	6	**5**	8	3
Ositettu perusmuotoh. (T5), AC	**3**	**2**	**19**	**6**	**2**	8	**6**
Ositettu perusmuotoh. (T5), ABC	3	2	19	6	**13**	8	6
Perinteinen (T1), ABCabc	**14**	**2**	**24**	**11**	**5**		
Finstems (T2), ACac	**14**	**2**	**23**	**11**	**2**		
Finstems (T2), ABCabc	**14**	**2**	**23**	**11**	**5**		
Hahmotin (T2), ACac	**14**	**2**	**23**	**11**	**2**		
Hahmotin (T2), ABCabc	**14**	**2**	**23**	**11**	**5**		
Seulonta (T3), ACac	**14**	**2**	**23**	**8**	**2**		
Seulonta (T3), ABCabc	**14**	**2**	**23**	**10**	**5**		
Perusmuotohak. (T4), Aa	**9**	**2**	**23**	**6**	**2**		
Perusmuotohak. (T4), ABab	**9**	**2**	**23**	**7**	**5**		
Perusmuotohak. (T4), ACac	**14**	**2**	**23**	**11**	**2**		
Perusmuotohak. (T4), ABCabc	**14**	**2**	**23**	**11**	**5**		
Ositettu perusm.h. (T5), Aa	**9**	**2**	**23**	**6**	**2**		
Ositettu perusm.h. (T5), ABab	**9**	**2**	**23**	**7**	**5**		
Ositettu perusm.h. (T5), ACac	**14**	**2**	**28**	**9**	**2**		**6**
Ositettu perusm.h. (T5), ABCabc	**14**	**2**	**28**	**11**	**14**		6

Tutkimusympäristö		**Y**			
ja kyselytyyppi	**25**	**26**	**27**	**28**	**29**
Perinteinen (T1), ABC	**2**	**2**	**9**	**3**	**5**
Finstems (T2), AC	**2**	**2**	**9**	**3**	**5**
Finstems (T2), ABC	2	2	9	3	5
Hahmotin (T2), AC	**2**	**2**	**9**	**3**	**5**
Hahmotin (T2), ABC	2	2	9	3	5
Seulonta (T3), AC	**2**	**2**	**8**	**0**	**5**
Seulonta (T3), ABC	2	2	8	0	5
Perusmuotohakemisto (T4), A	**2**	**2**	**8**	**2**	**5**
Perusmuotohakemisto (T4), AB	2	2	8	2	5
Perusmuotohakemisto (T4), AC	**2**	**2**	**9**	**3**	**5**
Perusmuotohakemisto (T4), ABC	2	2	9	3	5
Ositettu perusmuotoh. (T5), A	**2**	**2**	**8**	**2**	**5**
Ositettu perusmuotoh. (T5), AB	2	2	8	2	5
Ositettu perusmuotoh. (T5), AC	**2**	**2**	**9**	**3**	**5**
Ositettu perusmuotoh. (T5), ABC	2	2	9	3	5
Perinteinen (T1), ABCabc		**2**			**21**
Finstems (T2), ACac		**2**			**12**
Finstems (T2), ABCabc		2			**21**
Hahmotin (T2), ACac		**2**			**12**
Hahmotin (T2), ABCabc		2			**21**
Seulonta (T3), ACac		**2**			**7**
Seulonta (T3), ABCabc		2			**9**
Perusmuotohak. (T4), Aa		**2**			**7**
Perusmuotohak. (T4), ABab		2			**8**
Perusmuotohak. (T4), ACac		**2**			**12**
Perusmuotohak. (T4), ABCabc		2			**21**
Ositettu perusm.h. (T5), Aa		**2**			**7**
Ositettu perusm.h. (T5), ABab		2			**8**
Ositettu perusm.h. (T5), ACac		**2**	**10**	**3**	**7**
Ositettu perusm.h. (T5), ABCabc		2	10	3	**11**

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

Tutkimusympäristö ja kyselytyyppi	**N_p**	**Keski-arvo**	**Y N_y**	**Y Keski-arvo**	**J N_j**	**J Keski-arvo**
Perinteinen (T1), ABC	26	**11,3**	9	**6,1**	8	**15,9**
Finstems (T2), AC	26	8,3	9	5,8	8	6,3
Finstems (T2), ABC	26	**11,2**	9	**6,1**	8	**15,9**
Hahmotin (T2), AC	26	8,3	9	5,8	8	6,3
Hahmotin (T2), ABC	26	**11,0**	9	**6,1**	8	**15,1**
Seulonta (T3), AC	26	7,5	9	5,7	8	4,8
Seulonta (T3), ABC	26	**9,8**	9	**6,0**	8	**12,3**
Perusmuotohakemisto (T4), A	26	6,0	9	5,6	8	4,4
Perusmuotohakemisto (T4), AB	26	8,1	9	5,9	8	11,3
Perusmuotohakemisto (T4), AC	26	8,3	9	5,8	8	6,3
Perusmuotohakemisto (T4), ABC	26	**11,2**	9	**6,1**	8	**15,9**
Ositettu perusmuotoh. (T5), A	26	6,0	9	5,6	8	4,4
Ositettu perusmuotoh. (T5), AB	26	8,1	9	5,9	8	11,3
Ositettu perusmuotoh. (T5), AC	26	8,3	9	5,7	8	6,4
Ositettu perusmuotoh. (T5), ABC	26	**11,3**	9	**6,9**	8	**16,3**
Perinteinen (T1), ABCabc	14	**25,9**	9	**8,6**		
Finstems (T2), ACac	14	13,7	9	8,1		
Finstems (T2), ABCabc	14	**18,5**	9	**8,4**		
Hahmotin (T2), ACac	14	13,7	9	8,1		
Hahmotin (T2), ABCabc	14	**18,4**	9	**8,4**		
Seulonta (T3), ACac	14	11,6	9	7,8		
Seulonta (T3), ABCabc	14	**15,4**	9	**8,3**		
Perusmuotohak. (T4), Aa	14	7,7	9	6,8		
Perusmuotohak. (T4), ABab	14	11,2	9	7,2		
Perusmuotohak. (T4), ACac	14	12,1	9	8,1		
Perusmuotohak. (T4), ABCabc	14	**16,9**	9	**8,4**		
Ositettu perusm.h. (T5), Aa	14	7,7	9	6,8		
Ositettu perusm.h. (T5), ABab	14	11,2	9	7,2		
Ositettu perusm.h. (T5), ACac	14	14,3	9	8,6		
Ositettu perusm.h. (T5), ABCabc	14	**20,7**	9	**10,1**		

SAANNIN VAIHTELU

JA-operaattori

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

Tutkimusympäristö ja kyselytyyppi	**1**	**2**	**3**	**Y 4**	**Y 5**	**J 6**
Perint. (T1), ABC	**80 %**	**78 %**	**62 %**	**71 %**	**75 %**	**76 %**
Fins. (T2), AC	**80 %**	**78 %**	**62 %**	**71 %**	**75 %**	**35 %**
Fins. (T2), ABC	80 %	78 %	62 %	71 %	75 %	**76 %**
Hah. (T2), AC	**80 %**	**78 %**	**62 %**	**71 %**	**75 %**	**35 %**
Hah. (T2), ABC	80 %	78 %	62 %	71 %	75 %	**76 %**
Seul. (T3), AC	**80 %**	**78 %**	**62 %**	**71 %**	**75 %**	**35 %**
Seul. (T3), ABC	80 %	78 %	62 %	71 %	75 %	**53 %**
Perusm.h. (T4), A	**80 %**	**33 %**	**11 %**	**71 %**	**50 %**	**29 %**
Perusm.h. (T4), AB	80 %	33 %	11 %	71 %	50 %	**47 %**
Perusm.h. (T4), AC	**80 %**	**78 %**	**62 %**	**71 %**	**75 %**	**35 %**
Perusm.h. (T4), ABC	80 %	78 %	62 %	71 %	75 %	**76 %**
Osit. p.m.h. (T5), A	**80 %**	**33 %**	**11 %**	**71 %**	**50 %**	**29 %**
Osit. p.m.h. (T5), AB	80 %	33 %	11 %	71 %	50 %	**47 %**
Osit. p.m.h. (T5), AC	**80 %**	**78 %**	**62 %**	**71 %**	**75 %**	**35 %**
Osit. p.m.h. (T5), ABC	80 %	78 %	62 %	71 %	75 %	**88 %**
Perint. (T1), ABCabc			**64 %**	**88 %**	**100 %**	
Fins. (T2), ACac			**64 %**	**88 %**	**100 %**	
Fins. (T2), ABCabc			64 %	88 %	100 %	
Hah. (T2), ACac			**64 %**	**88 %**	**100 %**	
Hah. (T2), ABCabc			64 %	88 %	100 %	
Seul. (T3), ACac			**100 %**	**76 %**	**100 %**	
Seul. (T3), ABCabc			100 %	76 %	100 %	
Per.m.h. (T4), Aa			**30 %**	**71 %**	**75 %**	
Per.m.h. (T4), ABab			30 %	71 %	75 %	
Per.m.h. (T4), ACac			**100 %**	**88 %**	**100 %**	
Per.m.h. (T4), ABCabc			100 %	88 %	100 %	
Os. p.m.h. (T5), Aa			**30 %**	**71 %**	**75 %**	
Os.p.m.h. (T5), ABab			30 %	71 %	75 %	
Os.p.m.h. (T5), ACac			**100 %**	**100 %**	**100 %**	**47 %**
Os.p.m.h. (T5), ABCabc			100 %	100 %	100 %	**100 %**

= otanta

Tutkimusympäristö	**J**	**J**	**J**	**J**	**Y+J**	
ja kyselytyyppi	**7**	**8**	**9**	**10**	**11**	**12**
Perint. (T1), ABC	**96 %**	**100 %**	**100 %**	**100 %**	**100 %**	**100 %**
Fins. (T2), AC	**88 %**	**57 %**	**50 %**	**94 %**	**100 %**	**100 %**
Fins. (T2), ABC	**96 %**	**100 %**	**100 %**	**100 %**	**100 %**	100 %
Hah. (T2), AC	**88 %**	**57 %**	**50 %**	**94 %**	**100 %**	**100 %**
Hah. (T2), ABC	**96 %**	**86 %**	**100 %**	**94 %**	**100 %**	100 %
Seul. (T3), AC	**58 %**	**57 %**	**50 %**	**94 %**	**100 %**	**100 %**
Seul. (T3), ABC	**67 %**	**86 %**	**100 %**	**94 %**	**100 %**	100 %
Perusm.h. (T4), A	**79 %**	**57 %**	**50 %**	**72 %**	**100 %**	**50 %**
Perusm.h. (T4), AB	**88 %**	**86 %**	**100 %**	**89 %**	**100 %**	50 %
Perusm.h. (T4), AC	**88 %**	**57 %**	**50 %**	**94 %**	**100 %**	**100 %**
Perusm.h. (T4), ABC	**96 %**	**100 %**	**100 %**	**100 %**	**100 %**	100 %
Osit. p.m.h. (T5), A	**79 %**	**57 %**	**50 %**	**72 %**	**100 %**	**50 %**
Osit. p.m.h. (T5), AB	**88 %**	**86 %**	**100 %**	**89 %**	**100 %**	50 %
Osit. p.m.h. (T5), AC	**92 %**	**57 %**	**50 %**	**94 %**	**100 %**	**100 %**
Osit. p.m.h. (T5), ABC	**100 %**	**86 %**	**100 %**	**94 %**	**100 %**	100 %
Perint. (T1), ABCabc					**100 %**	**67 %**
Fins. (T2), ACac					**100 %**	**67 %**
Fins. (T2), ABCabc					**100 %**	**67 %**
Hah. (T2), ACac					**100 %**	**67 %**
Hah. (T2), ABCabc					**100 %**	**67 %**
Seul. (T3), ACac					**100 %**	**100 %**
Seul. (T3), ABCabc					**100 %**	**100 %**
Per.m.h. (T4), Aa					**100 %**	**50 %**
Per.m.h. (T4), ABab					**100 %**	**50 %**
Per.m.h. (T4), ACac					**100 %**	**67 %**
Per.m.h. (T4), ABCabc					**100 %**	**67 %**
Os. p.m.h. (T5), Aa					**100 %**	**50 %**
Os.p.m.h. (T5), ABab					**100 %**	**50 %**
Os.p.m.h. (T5), ACac	**92 %**				**100 %**	**50 %**
Os.p.m.h. (T5), ABCabc	**100 %**				**100 %**	**100 %**

Tutkimusympäristö	**J**		**Y**	**Y**	**Y**	**Y**
ja kyselytyyppi	**13**	**14**	**15**	**16**	**17**	**19**
Perint. (T1), ABC	**53 %**	**63 %**	**11 %**	**29 %**	**60 %**	**53 %**
Fins. (T2), AC	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**53 %**
Fins. (T2), ABC	**53 %**	60 %	11 %	29 %	60 %	53 %
Hah. (T2), AC	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**53 %**
Hah. (T2), ABC	**53 %**	60 %	11 %	29 %	60 %	53 %
Seul. (T3), AC	**11 %**	**53 %**	**11 %**	**29 %**	**60 %**	**42 %**
Seul. (T3), ABC	**53 %**	53 %	11 %	29 %	60 %	42 %
Perusm.h. (T4), A	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**37 %**
Perusm.h. (T4), AB	**50 %**	60 %	11 %	29 %	60 %	37 %
Perusm.h. (T4), AC	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**53 %**
Perusm.h. (T4), ABC	**53 %**	60 %	11 %	29 %	60 %	53 %
Osit. p.m.h. (T5), A	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**37 %**
Osit. p.m.h. (T5), AB	**50 %**	60 %	11 %	29 %	60 %	37 %
Osit. p.m.h. (T5), AC	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**42 %**
Osit. p.m.h. (T5), ABC	**53 %**	60 %	11 %	29 %	60 %	42 %
Perint. (T1), ABCabc	**100 %**	**97 %**	**100 %**	**71 %**	**96 %**	**95 %**
Fins. (T2), ACac	**33 %**	**73 %**	**100 %**	**43 %**	**96 %**	**95 %**
Fins. (T2), ABCabc	**100 %**	**100 %**	**100 %**	**71 %**	**96 %**	**95 %**
Hah. (T2), ACac	**33 %**	**73 %**	**100 %**	**43 %**	**96 %**	**95 %**
Hah. (T2), ABCabc	**100 %**	**93 %**	**100 %**	**71 %**	**96 %**	**95 %**
Seul. (T3), ACac	**33 %**	**70 %**	**100 %**	**43 %**	**88 %**	**63 %**
Seul. (T3), ABCabc	**89 %**	**73 %**	**100 %**	**57 %**	**92 %**	**84 %**
Per.m.h. (T4), Aa	**22 %**	**70 %**	**100 %**	**43 %**	**88 %**	**47 %**
Per.m.h. (T4), ABab	**75 %**	**90 %**	**100 %**	**57 %**	**88 %**	**63 %**
Per.m.h. (T4), ACac	**33 %**	**73 %**	**100 %**	**43 %**	**96 %**	**95 %**
Per.m.h. (T4), ABCabc	**100 %**	**93 %**	**100 %**	**71 %**	**96 %**	**95 %**
Os. p.m.h. (T5), Aa	**22 %**	**70 %**	**100 %**	**43 %**	**88 %**	**47 %**
Os.p.m.h. (T5), ABab	**75 %**	**90 %**	**100 %**	**57 %**	**88 %**	**63 %**
Os.p.m.h. (T5), ACac	**36 %**	**77 %**	**100 %**	**43 %**	**96 %**	**84 %**
Os.p.m.h. (T5), ABCabc	**100 %**	**83 %**	**100 %**	**100 %**	**96 %**	**89 %**

Tutkimusympäristö	**Y+J**				**Y**	
ja kyselytyyppi	**20**	**21**	**24**	**25**	**26**	**27**
Perint. (T1), ABC	**37 %**	**100 %**	**90 %**	**100 %**	**67 %**	**89 %**
Fins. (T2), AC	**21 %**	**100 %**	**90 %**	**100 %**	**67 %**	**89 %**
Fins. (T2), ABC	**37 %**	100 %	90 %	100 %	67 %	89 %
Hah. (T2), AC	**21 %**	**100 %**	**90 %**	**100 %**	**67 %**	**89 %**
Hah. (T2), ABC	**37 %**	100 %	90 %	100 %	67 %	89 %
Seul. (T3), AC	**21 %**	**100 %**	**90 %**	**100 %**	**67 %**	**89 %**
Seul. (T3), ABC	**37 %**	100 %	90 %	100 %	67 %	89 %
Perusm.h. (T4), A	**16 %**	**100 %**	**90 %**	**100 %**	**67 %**	**78 %**
Perusm.h. (T4), AB	**37 %**	100 %	90 %	100 %	67 %	78 %
Perusm.h. (T4), AC	**26 %**	100 %	**90 %**	**100 %**	**67 %**	**89 %**
Perusm.h. (T4), ABC	**42 %**	100 %	90 %	100 %	67 %	89 %
Osit. p.m.h. (T5), A	**16 %**	**100 %**	**90 %**	**100 %**	**67 %**	**78 %**
Osit. p.m.h. (T5), AB	**37 %**	100 %	90 %	100 %	67 %	78 %
Osit. p.m.h. (T5), AC	**26 %**	100 %	**90 %**	**100 %**	**67 %**	**100 %**
Osit. p.m.h. (T5), ABC	**89 %**	100 %	90 %	100 %	67 %	100 %
Perint. (T1), ABCabc	**42 %**				**100 %**	
Fins. (T2), ACac	**26 %**				**100 %**	
Fins. (T2), ABCabc	**42 %**				100 %	
Hah. (T2), ACac	**26 %**				**100 %**	
Hah. (T2), ABCabc	**42 %**				100 %	
Seul. (T3), ACac	**21 %**				**100 %**	
Seul. (T3), ABCabc	**42 %**				100 %	
Per.m.h. (T4), Aa	**16 %**				**67 %**	
Per.m.h. (T4), ABab	**37 %**				67 %	
Per.m.h. (T4), ACac	**32 %**				**100 %**	
Per.m.h. (T4), ABCabc	**47 %**				100 %	
Os. p.m.h. (T5), Aa	**16 %**				**67 %**	
Os.p.m.h. (T5), ABab	**37 %**				67 %	
Os.p.m.h. (T5), ACac	**37 %**		**100 %**		**100 %**	**100 %**
Os.p.m.h. (T5), ABCabc	**100 %**		100 %		100 %	100 %

Tutkimusympäristö ja kyselytyyppi	**28**	**29**
Perint. (T1), ABC	**27 %**	**56 %**
Fins. (T2), AC	**27 %**	**56 %**
Fins. (T2), ABC	27 %	56 %
Hah. (T2), AC	**27 %**	**56 %**
Hah. (T2), ABC	27 %	56 %
Seul. (T3), AC	**9 %**	**56 %**
Seul. (T3), ABC	9 %	56 %
Perusm.h. (T4), A	**27 %**	**56 %**
Perusm.h. (T4), AB	27 %	56 %
Perusm.h. (T4), AC	**27 %**	**56 %**
Perusm.h. (T4), ABC	27 %	56 %
Osit. p.m.h. (T5), A	**27 %**	**56 %**
Osit. p.m.h. (T5), AB	27 %	56 %
Osit. p.m.h. (T5), AC	**64 %**	**56 %**
Osit. p.m.h. (T5), ABC	64 %	56 %
Perint. (T1), ABCabc		**33 %**
Fins. (T2), ACac		**78 %**
Fins. (T2), ABCabc		**33 %**
Hah. (T2), ACac		**78 %**
Hah. (T2), ABCabc		**67 %**
Seul. (T3), ACac		**67 %**
Seul. (T3), ABCabc		**78 %**
Per.m.h. (T4), Aa		**67 %**
Per.m.h. (T4), ABab		**78 %**
Per.m.h. (T4), ACac		**78 %**
Per.m.h. (T4), ABCabc		**67 %**
Os. p.m.h. (T5), Aa		**67 %**
Os.p.m.h. (T5), ABab		**78 %**
Os.p.m.h. (T5), ACac	**100 %**	**78 %**
Os.p.m.h. (T5), ABCabc	100 %	**89 %**

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

			Y	Y	J	J
Tutkimusympäristö ja kyselytyyppi	**N_p**	**Keski-arvo**	**N_y**	**Keski-arvo**	**N_j**	**Keski-arvo**
Perint. (T1), ABC	26	**72,0 %**	9	**55,7 %**	8	**82,7 %**
Fins. (T2), AC	26	63,9 %	9	54,0 %	8	57,1 %
Fins. (T2), ABC	26	**71,8 %**	9	**55,7 %**	8	**82,7 %**
Hah. (T2), AC	26	63,9 %	9	54,0 %	8	57,1 %
Hah. (T2), ABC	26	**71,1 %**	9	**55,7 %**	8	**80,3 %**
Seul. (T3), AC	26	61,5 %	9	52,8 %	8	53,4 %
Seul. (T3), ABC	26	**67,7 %**	9	**54,5 %**	8	**73,7 %**
Perusm.h. (T4), A	26	54,7 %	9	48,8 %	8	51,9 %
Perusm.h. (T4), AB	26	61,7 %	9	51,2 %	8	74,5 %
Perusm.h. (T4), AC	26	64,1 %	9	54,5 %	8	57,7 %
Perusm.h. (T4), ABC	26	**72,0 %**	9	**56,3 %**	8	**83,4 %**
Osit. p.m.h. (T5), A	26	54,7 %	9	48,8 %	8	51,9 %
Osit. p.m.h. (T5), AB	26	61,7 %	9	51,2 %	8	74,5 %
Osit. p.m.h. (T5), AC	26	65,7 %	9	53,4 %	8	58,2 %
Osit. p.m.h. (T5), ABC	26	**75,1 %**	9	**60,4 %**	8	**88,8 %**
Perint. (T1), ABCabc	14	**82,4 %**	9	**88,1 %**		
Fins. (T2), ACac	14	75,9 %	9	83,1 %		
Fins. (T2), ABCabc	14	**82,6 %**	9	**88,1 %**		
Hah. (T2), ACac	14	75,9 %	9	83,1 %		
Hah. (T2), ABCabc	14	**84,5 %**	9	**88,1 %**		
Seul. (T3), ACac	14	75,8 %	9	76,8 %		
Seul. (T3), ABCabc	14	**85,1 %**	9	**83,5 %**		
Per.m.h. (T4), Aa	14	60,4 %	9	67,4 %		
Per.m.h. (T4), ABab	14	70,0 %	9	73,0 %		
Per.m.h. (T4), ACac	14	78,9 %	9	83,7 %		
Per.m.h. (T4), ABCabc	14	**87,5 %**	9	**88,6 %**		
Os. p.m.h. (T5), Aa	14	60,4 %	9	67,4 %		
Os.p.m.h. (T5), ABab	14	70,0 %	9	73,0 %		
Os.p.m.h. (T5), ACac	14	78,6 %	9	84,4 %		
Os.p.m.h. (T5), ABCabc	14	**97,0 %**	9	**98,4 %**		

SAANNIN VAIHTELU

virkeoperaattori

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

Tutkimusympäristö ja kyselytyyppi	**1**	**2**	**3**	**Y 4**	**Y 5**	**J 6**
Perint. (T1), ABC	**80 %**	**78 %**	**62 %**	**71 %**	**50 %**	**29 %**
Fins. (T2), AC	**80 %**	**78 %**	**62 %**	**71 %**	**50 %**	**12 %**
Fins. (T2), ABC	80 %	78 %	62 %	71 %	50 %	**29 %**
Hah. (T2), AC	**80 %**	**78 %**	**62 %**	**71 %**	**50 %**	**12 %**
Hah. (T2), ABC	80 %	78 %	62 %	71 %	50 %	**29 %**
Seul. (T3), AC	**80 %**	**78 %**	**62 %**	**71 %**	**50 %**	**12 %**
Seul. (T3), ABC	80 %	78 %	62 %	71 %	50 %	**24 %**
Perusm.h. (T4), A	**80 %**	**33 %**	**11 %**	**71 %**	**50 %**	**12 %**
Perusm.h. (T4), AB	80 %	33 %	11 %	71 %	50 %	**24 %**
Perusm.h. (T4), AC	**80 %**	**78 %**	**62 %**	**71 %**	**50 %**	**12 %**
Perusm.h. (T4), ABC	80 %	78 %	62 %	71 %	50 %	**29 %**
Osit. p.m.h. (T5), A	**80 %**	**33 %**	**11 %**	**71 %**	**50 %**	**12 %**
Osit. p.m.h. (T5), AB	80 %	33 %	11 %	71 %	50 %	**24 %**
Osit. p.m.h. (T5), AC	**80 %**	**78 %**	**62 %**	**71 %**	**50 %**	**12 %**
Osit. p.m.h. (T5), ABC	80 %	78 %	62 %	71 %	50 %	**29 %**
Perint. (T1), ABCabc			**79 %**	**76 %**	**75 %**	
Fins. (T2), ACac			**79 %**	**76 %**	**75 %**	
Fins. (T2), ABCabc			79 %	76 %	75 %	
Hah. (T2), ACac			**79 %**	**76 %**	**75 %**	
Hah. (T2), ABCabc			79 %	76 %	75 %	
Seul. (T3), ACac			**72 %**	**76 %**	**75 %**	
Seul. (T3), ABCabc			72 %	76 %	75 %	
Per.m.h. (T4), Aa			**15 %**	**71 %**	**75 %**	
Per.m.h. (T4), ABab			15 %	71 %	75 %	
Per.m.h. (T4), ACac			**45 %**	**76 %**	**75 %**	
Per.m.h. (T4), ABCabc			45 %	76 %	75 %	
Os. p.m.h. (T5), Aa			**15 %**	**71 %**	**75 %**	
Os.p.m.h. (T5), ABab			15 %	71 %	75 %	
Os.p.m.h. (T5), ACac			**77 %**	**76 %**	**100 %**	**12 %**
Os.p.m.h. (T5), ABCabc			77 %	76 %	100 %	**35 %**

= otanta

Tutkimusympäristö ja kyselytyyppi	**J 7**	**J 8**	**J 9**	**J 10**	**Y+J 11**	**12**
Perint. (T1), ABC	**42 %**	**71 %**	**50 %**	**83 %**	**100 %**	**100 %**
Fins. (T2), AC	**33 %**	**43 %**	**33 %**	**61 %**	**100 %**	**100 %**
Fins. (T2), ABC	**42 %**	**71 %**	**50 %**	**83 %**	**100 %**	100 %
Hah. (T2), AC	**33 %**	**43 %**	**33 %**	**61 %**	**100 %**	**100 %**
Hah. (T2), ABC	**42 %**	**57 %**	**50 %**	**78 %**	**100 %**	100 %
Seul. (T3), AC	**4 %**	**43 %**	**33 %**	**61 %**	**100 %**	**100 %**
Seul. (T3), ABC	**17 %**	**57 %**	**50 %**	**78 %**	**100 %**	100 %
Perusm.h. (T4), A	**29 %**	**29 %**	**33 %**	**50 %**	**100 %**	**50 %**
Perusm.h. (T4), AB	**38 %**	**57 %**	**50 %**	**78 %**	**100 %**	50 %
Perusm.h. (T4), AC	**33 %**	**43 %**	**33 %**	**61 %**	**100 %**	**100 %**
Perusm.h. (T4), ABC	**42 %**	**71 %**	**50 %**	**83 %**	**100 %**	100 %
Osit. p.m.h. (T5), A	**29 %**	**29 %**	**33 %**	**50 %**	**100 %**	**50 %**
Osit. p.m.h. (T5), AB	**38 %**	**57 %**	**50 %**	**78 %**	**100 %**	50 %
Osit. p.m.h. (T5), AC	**33 %**	**43 %**	**33 %**	**61 %**	**100 %**	**100 %**
Osit. p.m.h. (T5), ABC	**42 %**	**57 %**	**50 %**	**78 %**	**100 %**	100 %
Perint. (T1), ABCabc					**100 %**	**100 %**
Fins. (T2), ACac					**100 %**	**100 %**
Fins. (T2), ABCabc					**100 %**	**100 %**
Hah. (T2), ACac					**100 %**	**100 %**
Hah. (T2), ABCabc					**100 %**	**100 %**
Seul. (T3), ACac					**100 %**	**100 %**
Seul. (T3), ABCabc					**100 %**	**100 %**
Per.m.h. (T4), Aa					**100 %**	**50 %**
Per.m.h. (T4), ABab					**100 %**	**50 %**
Per.m.h. (T4), ACac					**100 %**	**100 %**
Per.m.h. (T4), ABCabc					**100 %**	**100 %**
Os. p.m.h. (T5), Aa					**100 %**	**50 %**
Os.p.m.h. (T5), ABab					**100 %**	**50 %**
Os.p.m.h. (T5), ACac	**33 %**				**100 %**	**100 %**
Os.p.m.h. (T5), ABCabc	**42 %**				**100 %**	**100 %**

Tutkimusympäristö	**J**		**Y**	**Y**	**Y**	**Y**
ja kyselytyyppi	**13**	**14**	**15**	**16**	**17**	**19**
Perint. (T1), ABC	**53 %**	**63 %**	**11 %**	**29 %**	**60 %**	**32 %**
Fins. (T2), AC	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**32 %**
Fins. (T2), ABC	**53 %**	60 %	11 %	29 %	60 %	32 %
Hah. (T2), AC	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**32 %**
Hah. (T2), ABC	**53 %**	60 %	11 %	29 %	60 %	32 %
Seul. (T3), AC	**11 %**	**53 %**	**11 %**	**29 %**	**60 %**	**26 %**
Seul. (T3), ABC	**53 %**	53 %	11 %	29 %	60 %	26 %
Perusm.h. (T4), A	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**26 %**
Perusm.h. (T4), AB	**50 %**	60 %	11 %	29 %	60 %	26 %
Perusm.h. (T4), AC	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**32 %**
Perusm.h. (T4), ABC	**53 %**	60 %	11 %	29 %	60 %	32 %
Osit. p.m.h. (T5), A	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**26 %**
Osit. p.m.h. (T5), AB	**50 %**	60 %	11 %	29 %	60 %	26 %
Osit. p.m.h. (T5), AC	**11 %**	**60 %**	**11 %**	**29 %**	**60 %**	**26 %**
Osit. p.m.h. (T5), ABC	**53 %**	60 %	11 %	29 %	60 %	26 %
Perint. (T1), ABCabc	**78 %**	**67 %**	**78 %**	**29 %**	**72 %**	**47 %**
Fins. (T2), ACac	**22 %**	**63 %**	**78 %**	**29 %**	**72 %**	**47 %**
Fins. (T2), ABCabc	**78 %**	**73 %**	**78 %**	**29 %**	**72 %**	**47 %**
Hah. (T2), ACac	**22 %**	**63 %**	**78 %**	**29 %**	**72 %**	**47 %**
Hah. (T2), ABCabc	**78 %**	**73 %**	**78 %**	**29 %**	**72 %**	**47 %**
Seul. (T3), ACac	**17 %**	**57 %**	**78 %**	**29 %**	**72 %**	**32 %**
Seul. (T3), ABCabc	**75 %**	**60 %**	**78 %**	**29 %**	**72 %**	**42 %**
Per.m.h. (T4), Aa	**11 %**	**60 %**	**56 %**	**29 %**	**72 %**	**26 %**
Per.m.h. (T4), ABab	**67 %**	**67 %**	**56 %**	**29 %**	**72 %**	**32 %**
Per.m.h. (T4), ACac	**22 %**	**63 %**	**78 %**	**29 %**	**72 %**	**47 %**
Per.m.h. (T4), ABCabc	**78 %**	**73 %**	**78 %**	**29 %**	**72 %**	**47 %**
Os. p.m.h. (T5), Aa	**11 %**	**60 %**	**56 %**	**29 %**	**72 %**	**26 %**
Os.p.m.h. (T5), ABab	**67 %**	**67 %**	**56 %**	**29 %**	**72 %**	**32 %**
Os.p.m.h. (T5), ACac	**22 %**	**70 %**	**78 %**	**29 %**	**80 %**	**37 %**
Os.p.m.h. (T5), ABCabc	**94 %**	**80 %**	**78 %**	**29 %**	**80 %**	**47 %**

Tutkimusympäristö	**Y+J**				**Y**	
ja kyselytyyppi	**20**	**21**	**24**	**25**	**26**	**27**
Perint. (T1), ABC	**21 %**	**100 %**	**60 %**	**100 %**	**67 %**	**78 %**
Fins. (T2), AC	**5 %**	**100 %**	**60 %**	**100 %**	**67 %**	**78 %**
Fins. (T2), ABC	**21 %**	100 %	60 %	100 %	67 %	78 %
Hah. (T2), AC	**5 %**	**100 %**	**60 %**	**100 %**	**67 %**	**78 %**
Hah. (T2), ABC	**21 %**	100 %	60 %	100 %	67 %	78 %
Seul. (T3), AC	**5 %**	**100 %**	**60 %**	**100 %**	**67 %**	**78 %**
Seul. (T3), ABC	**21 %**	100 %	60 %	100 %	67 %	78 %
Perusm.h. (T4), A	**11 %**	**100 %**	**30 %**	**100 %**	**67 %**	**78 %**
Perusm.h. (T4), AB	**26 %**	100 %	30 %	100 %	67 %	78 %
Perusm.h. (T4), AC	**11 %**	100 %	**60 %**	**100 %**	**67 %**	**78 %**
Perusm.h. (T4), ABC	**26 %**	100 %	60 %	100 %	67 %	78 %
Osit. p.m.h. (T5), A	**11 %**	**100 %**	**30 %**	**100 %**	**67 %**	**78 %**
Osit. p.m.h. (T5), AB	**26 %**	100 %	30 %	100 %	67 %	78 %
Osit. p.m.h. (T5), AC	**11 %**	100 %	**60 %**	**100 %**	**67 %**	**89 %**
Osit. p.m.h. (T5), ABC	**63 %**	100 %	60 %	100 %	67 %	89 %
Perint. (T1), ABCabc	**21 %**				**67 %**	
Fins. (T2), ACac	**5 %**				**67 %**	
Fins. (T2), ABCabc	**21 %**				67 %	
Hah. (T2), ACac	**5 %**				**67 %**	
Hah. (T2), ABCabc	**21 %**				67 %	
Seul. (T3), ACac	**5 %**				**67 %**	
Seul. (T3), ABCabc	**21 %**				67 %	
Per.m.h. (T4), Aa	**11 %**				**67 %**	
Per.m.h. (T4), ABab	**26 %**				67 %	
Per.m.h. (T4), ACac	**11 %**				**67 %**	
Per.m.h. (T4), ABCabc	**26 %**				67 %	
Os. p.m.h. (T5), Aa	**11 %**				**67 %**	
Os.p.m.h. (T5), ABab	**26 %**				67 %	
Os.p.m.h. (T5), ACac	**11 %**		**60 %**		**67 %**	**89 %**
Os.p.m.h. (T5), ABCabc	**68 %**		60 %		67 %	89 %

Tutkimusympäristö ja kyselytyyppi	**28**	**29**
Perint. (T1), ABC	**18 %**	**56 %**
Fins. (T2), AC	**18 %**	**56 %**
Fins. (T2), ABC	18 %	56 %
Hah. (T2), AC	**18 %**	**56 %**
Hah. (T2), ABC	18 %	56 %
Seul. (T3), AC	**0 %**	**56 %**
Seul. (T3), ABC	0 %	56 %
Perusm.h. (T4), A	**18 %**	**56 %**
Perusm.h. (T4), AB	18 %	56 %
Perusm.h. (T4), AC	**18 %**	**56 %**
Perusm.h. (T4), ABC	18 %	56 %
Osit. p.m.h. (T5), A	**18 %**	**56 %**
Osit. p.m.h. (T5), AB	18 %	56 %
Osit. p.m.h. (T5), AC	**18 %**	**56 %**
Osit. p.m.h. (T5), ABC	18 %	56 %
Perint. (T1), ABCabc		**67 %**
Fins. (T2), ACac		**67 %**
Fins. (T2), ABCabc		**67 %**
Hah. (T2), ACac		**67 %**
Hah. (T2), ABCabc		**67 %**
Seul. (T3), ACac		**67 %**
Seul. (T3), ABCabc		**67 %**
Per.m.h. (T4), Aa		**67 %**
Per.m.h. (T4), ABab		**67 %**
Per.m.h. (T4), ACac		**67 %**
Per.m.h. (T4), ABCabc		**67 %**
Os. p.m.h. (T5), Aa		**67 %**
Os.p.m.h. (T5), ABab		**67 %**
Os.p.m.h. (T5), ACac	**18 %**	**67 %**
Os.p.m.h. (T5), ABCabc	18 %	**78 %**

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

			Y	Y	J	J
Tutkimusympäristö ja kyselytyyppi	**N_p**	**Keski-arvo**	**N_y**	**Keski-arvo**	**N_j**	**Keski-arvo**
Perint. (T1), ABC	26	**60,1 %**	9	**48,8 %**	8	**56,2 %**
Fins. (T2), AC	26	54,2 %	9	47,1 %	8	37,3 %
Fins. (T2), ABC	26	**60,0 %**	9	**48,8 %**	8	**56,2 %**
Hah. (T2), AC	26	54,2 %	9	47,1 %	8	37,3 %
Hah. (T2), ABC	26	**59,2 %**	9	**48,8 %**	8	**53,7 %**
Seul. (T3), AC	26	51,9 %	9	46,5 %	8	33,7 %
Seul. (T3), ABC	26	**56,9 %**	9	**48,3 %**	8	**49,9 %**
Perusm.h. (T4), A	26	46,3 %	9	47,1 %	8	34,3 %
Perusm.h. (T4), AB	26	52,0 %	9	48,8 %	8	52,8 %
Perusm.h. (T4), AC	26	54,4 %	9	47,7 %	8	38,0 %
Perusm.h. (T4), ABC	26	**60,2 %**	9	**49,4 %**	8	**56,9 %**
Osit. p.m.h. (T5), A	26	46,3 %	9	47,1 %	8	34,3 %
Osit. p.m.h. (T5), AB	26	52,0 %	9	48,8 %	8	52,8 %
Osit. p.m.h. (T5), AC	26	54,6 %	9	47,1 %	8	38,0 %
Osit. p.m.h. (T5), ABC	26	**61,0 %**	9	**52,9 %**	8	**59,0 %**
Perint. (T1), ABCabc	14	**68,2 %**	9	**62,8 %**		
Fins. (T2), ACac	14	62,9 %	9	61,0 %		
Fins. (T2), ABCabc	14	**68,7 %**	9	**62,8 %**		
Hah. (T2), ACac	14	62,9 %	9	61,0 %		
Hah. (T2), ABCabc	14	**68,7 %**	9	**62,8 %**		
Seul. (T3), ACac	14	60,4 %	9	59,3 %		
Seul. (T3), ABCabc	14	**66,7 %**	9	**62,2 %**		
Per.m.h. (T4), Aa	14	50,6 %	9	56,1 %		
Per.m.h. (T4), ABab	14	56,5 %	9	58,5 %		
Per.m.h. (T4), ACac	14	60,8 %	9	61,6 %		
Per.m.h. (T4), ABCabc	14	**66,6 %**	9	**63,4 %**		
Os. p.m.h. (T5), Aa	14	50,6 %	9	56,1 %		
Os.p.m.h. (T5), ABab	14	56,5 %	9	58,5 %		
Os.p.m.h. (T5), ACac	14	65,2 %	9	64,1 %		
Os.p.m.h. (T5), ABCabc	14	**76,7 %**	9	**71,7 %**		

TARKKUUDEN VAIHTELU

JA-operaattori

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

Tutkimusympäristö ja kyselytyyppi	**1**	**2**	**3**	**Y 4**	**Y 5**	**J 6**
Perint. (T1), ABC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**50 %**
Fins. (T2), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**40 %**
Fins. (T2), ABC	50 %	78 %	74 %	92 %	100 %	**50 %**
Hah. (T2), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**40 %**
Hah. (T2), ABC	50 %	78 %	74 %	92 %	100 %	**50 %**
Seul. (T3), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**43 %**
Seul. (T3), ABC	50 %	78 %	74 %	92 %	100 %	**45 %**
Perusm.h. (T4), A	**57 %**	**60 %**	**63 %**	**92 %**	**100 %**	**45 %**
Perusm.h. (T4), AB	57 %	60 %	63 %	92 %	100 %	**47 %**
Perusm.h. (T4), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**40 %**
Perusm.h. (T4), ABC	50 %	78 %	74 %	92 %	100 %	**50 %**
Osit. p.m.h. (T5), A	**57 %**	**60 %**	**63 %**	**92 %**	**100 %**	**45 %**
Osit. p.m.h. (T5), AB	57 %	60 %	63 %	92 %	100 %	**47 %**
Osit. p.m.h. (T5), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**38 %**
Osit. p.m.h. (T5), ABC	50 %	78 %	74 %	92 %	100 %	**52 %**
Perint. (T1), ABCabc			**20 %**	**83 %**	**25 %**	
Fins. (T2), ACac			**20 %**	**83 %**	**25 %**	
Fins. (T2), ABCabc			20 %	83 %	25 %	
Hah. (T2), ACac			**20 %**	**83 %**	**25 %**	
Hah. (T2), ABCabc			20 %	83 %	25 %	
Seul. (T3), ACac			**52 %**	**93 %**	**27 %**	
Seul. (T3), ABCabc			52 %	93 %	27 %	
Per.m.h. (T4), Aa			**36 %**	**80 %**	**33 %**	
Per.m.h. (T4), ABab			36 %	80 %	33 %	
Per.m.h. (T4), ACac			**36 %**	**83 %**	**25 %**	
Per.m.h. (T4), ABCabc			36 %	83 %	25 %	
Os. p.m.h. (T5), Aa			**36 %**	**80 %**	**33 %**	
Os.p.m.h. (T5), ABab			36 %	80 %	33 %	
Os.p.m.h. (T5), ACac			**48 %**	**77 %**	**19 %**	**17 %**
Os.p.m.h. (T5), ABCabc			48 %	77 %	19 %	**22 %**

= otanta

Tutkimusympäristö	**J**	**J**	**J**	**J**	**Y+J**	
ja kyselytyyppi	**7**	**8**	**9**	**10**	**11**	**12**
Perint. (T1), ABC	**70 %**	**5 %**	**13 %**	**46 %**	**50 %**	**75 %**
Fins. (T2), AC	**72 %**	**8 %**	**50 %**	**57 %**	**50 %**	**75 %**
Fins. (T2), ABC	**70 %**	**5 %**	**13 %**	**46 %**	**50 %**	75 %
Hah. (T2), AC	**72 %**	**8 %**	**50 %**	**57 %**	**50 %**	**75 %**
Hah. (T2), ABC	**70 %**	**5 %**	**13 %**	**45 %**	**50 %**	75 %
Seul. (T3), AC	**88 %**	**15 %**	**50 %**	**57 %**	**50 %**	**75 %**
Seul. (T3), ABC	**84 %**	**7 %**	**15 %**	**45 %**	**50 %**	75 %
Perusm.h. (T4), A	**76 %**	**15 %**	**60 %**	**62 %**	**100 %**	**60 %**
Perusm.h. (T4), AB	**72 %**	**8 %**	**24 %**	**48 %**	**100 %**	60 %
Perusm.h. (T4), AC	**72 %**	**8 %**	**50 %**	**57 %**	**50 %**	**75 %**
Perusm.h. (T4), ABC	**70 %**	**5 %**	**13 %**	**46 %**	**50 %**	75 %
Osit. p.m.h. (T5), A	**76 %**	**15 %**	**60 %**	**62 %**	**100 %**	**60 %**
Osit. p.m.h. (T5), AB	**72 %**	**8 %**	**24 %**	**48 %**	**100 %**	60 %
Osit. p.m.h. (T5), AC	**73 %**	**8 %**	**33 %**	**57 %**	**50 %**	**75 %**
Osit. p.m.h. (T5), ABC	**71 %**	**5 %**	**12 %**	**45 %**	**50 %**	75 %
Perint. (T1), ABCabc					**25 %**	**4 %**
Fins. (T2), ACac					**50 %**	**3 %**
Fins. (T2), ABCabc					**25 %**	**3 %**
Hah. (T2), ACac					**50 %**	**3 %**
Hah. (T2), ABCabc					**25 %**	**3 %**
Seul. (T3), ACac					**50 %**	**13 %**
Seul. (T3), ABCabc					**25 %**	**8 %**
Per.m.h. (T4), Aa					**100 %**	**8 %**
Per.m.h. (T4), ABab					**33 %**	**6 %**
Per.m.h. (T4), ACac					**50 %**	**3 %**
Per.m.h. (T4), ABCabc					**25 %**	**3 %**
Os. p.m.h. (T5), Aa					**100 %**	**8 %**
Os.p.m.h. (T5), ABab					**33 %**	**6 %**
Os.p.m.h. (T5), ACac	**51 %**				**50 %**	**3 %**
Os.p.m.h. (T5), ABCabc	**50 %**				**25 %**	**15 %**

Tutkimusympäristö	**J**		**Y**	**Y**	**Y**	**Y**
ja kyselytyyppi	**13**	**14**	**15**	**16**	**17**	**19**
Perint. (T1), ABC	**63 %**	**90 %**	**33 %**	**100 %**	**79 %**	**53 %**
Fins. (T2), AC	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**71 %**
Fins. (T2), ABC	**63 %**	90 %	33 %	100 %	79 %	71 %
Hah. (T2), AC	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**71 %**
Hah. (T2), ABC	**63 %**	90 %	33 %	100 %	79 %	71 %
Seul. (T3), AC	**80 %**	**89 %**	**33 %**	**100 %**	**79 %**	**67 %**
Seul. (T3), ABC	**63 %**	**89 %**	33 %	100 %	79 %	67 %
Perusm.h. (T4), A	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**70 %**
Perusm.h. (T4), AB	**64 %**	90 %	33 %	100 %	79 %	70 %
Perusm.h. (T4), AC	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**71 %**
Perusm.h. (T4), ABC	**63 %**	90 %	33 %	100 %	79 %	71 %
Osit. p.m.h. (T5), A	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**70 %**
Osit. p.m.h. (T5), AB	**64 %**	90 %	33 %	100 %	79 %	70 %
Osit. p.m.h. (T5), AC	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**67 %**
Osit. p.m.h. (T5), ABC	**63 %**	90 %	33 %	100 %	79 %	67 %
Perint. (T1), ABCabc	**24 %**	**4 %**	**35 %**	**11 %**	**33 %**	**56 %**
Fins. (T2), ACac	**28 %**	**58 %**	**35 %**	**13 %**	**35 %**	**62 %**
Fins. (T2), ABCabc	**24 %**	**35 %**	**35 %**	**11 %**	**35 %**	**62 %**
Hah. (T2), ACac	**28 %**	**58 %**	**35 %**	**13 %**	**35 %**	**62 %**
Hah. (T2), ABCabc	**24 %**	**27 %**	**35 %**	**11 %**	**35 %**	**62 %**
Seul. (T3), ACac	**38 %**	**88 %**	**41 %**	**23 %**	**56 %**	**55 %**
Seul. (T3), ABCabc	**29 %**	**51 %**	**41 %**	**20 %**	**49 %**	**59 %**
Per.m.h. (T4), Aa	**33 %**	**88 %**	**50 %**	**27 %**	**65 %**	**69 %**
Per.m.h. (T4), ABab	**34 %**	**52 %**	**50 %**	**22 %**	**56 %**	**75 %**
Per.m.h. (T4), ACac	**28 %**	**71 %**	**35 %**	**13 %**	**35 %**	**62 %**
Per.m.h. (T4), ABCabc	**24 %**	**28 %**	**35 %**	**11 %**	**35 %**	**62 %**
Os. p.m.h. (T5), Aa	**33 %**	**88 %**	**50 %**	**27 %**	**65 %**	**69 %**
Os.p.m.h. (T5), ABab	**34 %**	**52 %**	**50 %**	**22 %**	**56 %**	**75 %**
Os.p.m.h. (T5), ACac	**33 %**	**33 %**	**36 %**	**16 %**	**41 %**	**62 %**
Os.p.m.h. (T5), ABCabc	**31 %**	**19 %**	**36 %**	**21 %**	**38 %**	**57 %**

Tutkimusympäristö	**Y+J**				**Y**	
ja kyselytyyppi	**20**	**21**	**24**	**25**	**26**	**27**
Perint. (T1), ABC	**78 %**	**55 %**	**82 %**	**100 %**	**100 %**	**73 %**
Fins. (T2), AC	**80 %**	**55 %**	**82 %**	**100 %**	**100 %**	**73 %**
Fins. (T2), ABC	**78 %**	55 %	82 %	100 %	100 %	73 %
Hah. (T2), AC	**80 %**	**55 %**	**82 %**	**100 %**	**100 %**	**73 %**
Hah. (T2), ABC	**78 %**	55 %	82 %	100 %	100 %	73 %
Seul. (T3), AC	**80 %**	**60 %**	**82 %**	**100 %**	**100 %**	**80 %**
Seul. (T3), ABC	**78 %**	60 %	82 %	100 %	100 %	80 %
Perusm.h. (T4), A	**100 %**	**75 %**	**90 %**	**100 %**	**100 %**	**88 %**
Perusm.h. (T4), AB	**88 %**	75 %	90 %	100 %	100 %	88 %
Perusm.h. (T4), AC	**100 %**	75 %	**90 %**	**100 %**	**100 %**	**73 %**
Perusm.h. (T4), ABC	**89 %**	75 %	90 %	100 %	100 %	73 %
Osit. p.m.h. (T5), A	**100 %**	**75 %**	**90 %**	**100 %**	**100 %**	**88 %**
Osit. p.m.h. (T5), AB	**88 %**	75 %	90 %	100 %	100 %	88 %
Osit. p.m.h. (T5), AC	**100 %**	75 %	**90 %**	**100 %**	**100 %**	**82 %**
Osit. p.m.h. (T5), ABC	**85 %**	75 %	90 %	100 %	100 %	82 %
Perint. (T1), ABCabc	**73 %**				**38 %**	
Fins. (T2), ACac	**83 %**				**38 %**	
Fins. (T2), ABCabc	**73 %**				38 %	
Hah. (T2), ACac	**83 %**				**38 %**	
Hah. (T2), ABCabc	**80 %**				38 %	
Seul. (T3), ACac	**80 %**				**60 %**	
Seul. (T3), ABCabc	**80 %**				60 %	
Per.m.h. (T4), Aa	**100 %**				**67 %**	
Per.m.h. (T4), ABab	**88 %**				67 %	
Per.m.h. (T4), ACac	**100 %**				**43 %**	
Per.m.h. (T4), ABCabc	**82 %**				43 %	
Os. p.m.h. (T5), Aa	**100 %**				**67 %**	
Os.p.m.h. (T5), ABab	**88 %**				67 %	
Os.p.m.h. (T5), ACac	**100 %**		**63 %**		**60 %**	**75 %**
Os.p.m.h. (T5), ABCabc	**86 %**		63 %		60 %	75 %

Tutkimusympäristö ja kyselytyyppi	**28**	**29**
Perint. (T1), ABC	**60 %**	**100 %**
Fins. (T2), AC	**60 %**	**100 %**
Fins. (T2), ABC	60 %	100 %
Hah. (T2), AC	**60 %**	**100 %**
Hah. (T2), ABC	60 %	100 %
Seul. (T3), AC	**100 %**	**100 %**
Seul. (T3), ABC	100 %	100 %
Perusm.h. (T4), A	**75 %**	**100 %**
Perusm.h. (T4), AB	75 %	100 %
Perusm.h. (T4), AC	**60 %**	**100 %**
Perusm.h. (T4), ABC	60 %	100 %
Osit. p.m.h. (T5), A	**75 %**	**100 %**
Osit. p.m.h. (T5), AB	75 %	100 %
Osit. p.m.h. (T5), AC	**78 %**	**100 %**
Osit. p.m.h. (T5), ABC	78 %	100 %
Perint. (T1), ABCabc		**3 %**
Fins. (T2), ACac		**21 %**
Fins. (T2), ABCabc		**3 %**
Hah. (T2), ACac		**21 %**
Hah. (T2), ABCabc		**7 %**
Seul. (T3), ACac		**43 %**
Seul. (T3), ABCabc		**24 %**
Per.m.h. (T4), Aa		**46 %**
Per.m.h. (T4), ABab		**25 %**
Per.m.h. (T4), ACac		**21 %**
Per.m.h. (T4), ABCabc		**7 %**
Os. p.m.h. (T5), Aa		**46 %**
Os.p.m.h. (T5), ABab		**25 %**
Os.p.m.h. (T5), ACac	**69 %**	**35 %**
Os.p.m.h. (T5), ABCabc	69 %	**21 %**

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

			Y	Y	J	J
Tutkimusympäristö ja kyselytyyppi	**N_p**	**Keski-arvo**	**N_y**	**Keski-arvo**	**N_j**	**Keski-arvo**
Perint. (T1), ABC	26	**68,0 %**	9	**76,1 %**	8	**46,9 %**
Fins. (T2), AC	26	71,1 %	9	78,4 %	8	54,7 %
Fins. (T2), ABC	26	**68,7 %**	9	**78,2 %**	8	**46,9 %**
Hah. (T2), AC	26	71,1 %	9	78,4 %	8	54,7 %
Hah. (T2), ABC	26	**68,7 %**	9	**78,2 %**	8	**46,7 %**
Seul. (T3), AC	26	73,9 %	9	77,9 %	8	57,7 %
Seul. (T3), ABC	26	**71,0 %**	9	**77,7 %**	8	**48,3 %**
Perusm.h. (T4), A	26	75,8 %	9	86,1 %	8	67,3 %
Perusm.h. (T4), AB	26	72,4 %	9	84,7 %	8	56,5 %
Perusm.h. (T4), AC	26	73,0 %	9	80,7 %	8	57,2 %
Perusm.h. (T4), ABC	26	**70,3 %**	9	**79,4 %**	8	**48,2 %**
Osit. p.m.h. (T5), A	26	75,8 %	9	86,1 %	8	67,3 %
Osit. p.m.h. (T5), AB	26	72,4 %	9	84,7 %	8	56,5 %
Osit. p.m.h. (T5), AC	26	73,2 %	9	80,1 %	8	54,9 %
Osit. p.m.h. (T5), ABC	26	**71,0 %**	9	**78,5 %**	8	**47,8 %**
Perint. (T1), ABCabc	14	**31,0 %**	9	**42,1 %**		
Fins. (T2), ACac	14	39,5 %	9	47,1 %		
Fins. (T2), ABCabc	14	**33,7 %**	9	**43,0 %**		
Hah. (T2), ACac	14	39,5 %	9	47,0 %		
Hah. (T2), ABCabc	14	**33,9 %**	9	**43,7 %**		
Seul. (T3), ACac	14	51,3 %	9	53,8 %		
Seul. (T3), ABCabc	14	**44,1 %**	9	**50,4 %**		
Per.m.h. (T4), Aa	14	57,3 %	9	65,7 %		
Per.m.h. (T4), ABab	14	47,0 %	9	56,1 %		
Per.m.h. (T4), ACac	14	43,2 %	9	49,5 %		
Per.m.h. (T4), ABCabc	14	**35,7 %**	9	**44,6 %**		
Os. p.m.h. (T5), Aa	14	57,3 %	9	65,7 %		
Os.p.m.h. (T5), ABab	14	47,0 %	9	56,1 %		
Os.p.m.h. (T5), ACac	14	43,8 %	9	51,1 %		
Os.p.m.h. (T5), ABCabc	14	**39,5 %**	9	**46,6 %**		

TARKKUUDEN VAIHTELU

virkeoperaattori

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

Tutkimusympäristö ja kyselytyyppi	**1**	**2**	**3**	**Y 4**	**Y 5**	**J 6**
Perint. (T1), ABC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**71 %**
Fins. (T2), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**67 %**
Fins. (T2), ABC	50 %	78 %	74 %	92 %	100 %	**71 %**
Hah. (T2), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**67 %**
Hah. (T2), ABC	50 %	78 %	74 %	92 %	100 %	**71 %**
Seul. (T3), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**67 %**
Seul. (T3), ABC	50 %	78 %	74 %	92 %	100 %	**80 %**
Perusm.h. (T4), A	**57 %**	**60 %**	**63 %**	**92 %**	**100 %**	**67 %**
Perusm.h. (T4), AB	57 %	60 %	63 %	92 %	100 %	**67 %**
Perusm.h. (T4), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**67 %**
Perusm.h. (T4), ABC	50 %	78 %	74 %	92 %	100 %	**71 %**
Osit. p.m.h. (T5), A	**57 %**	**60 %**	**63 %**	**92 %**	**100 %**	**67 %**
Osit. p.m.h. (T5), AB	57 %	60 %	63 %	92 %	100 %	**67 %**
Osit. p.m.h. (T5), AC	**50 %**	**78 %**	**74 %**	**92 %**	**100 %**	**67 %**
Osit. p.m.h. (T5), ABC	50 %	78 %	74 %	92 %	100 %	**71 %**
Perint. (T1), ABCabc			**66 %**	**93 %**	**100 %**	
Fins. (T2), ACac			**66 %**	**93 %**	**100 %**	
Fins. (T2), ABCabc			66 %	93 %	100 %	
Hah. (T2), ACac			**66 %**	**93 %**	**100 %**	
Hah. (T2), ABCabc			66 %	93 %	100 %	
Seul. (T3), ACac			**71 %**	**93 %**	**100 %**	
Seul. (T3), ABCabc			71 %	93 %	100 %	
Per.m.h. (T4), Aa			**70 %**	**92 %**	**100 %**	
Per.m.h. (T4), ABab			70 %	92 %	100 %	
Per.m.h. (T4), ACac			**62 %**	**93 %**	**100 %**	
Per.m.h. (T4), ABCabc			62 %	93 %	100 %	
Os. p.m.h. (T5), Aa			**70 %**	**92 %**	**100 %**	
Os.p.m.h. (T5), ABab			70 %	92 %	100 %	
Os.p.m.h. (T5), ACac			**64 %**	**93 %**	**100 %**	**67 %**
Os.p.m.h. (T5), ABCabc			64 %	93 %	100 %	**75 %**

= otanta

Tutkimusympäristö ja kyselytyyppi	**J 7**	**J 8**	**J 9**	**J 10**	**Y+J 11**	**12**
Perint. (T1), ABC	**83 %**	**13 %**	**100 %**	**54 %**	**50 %**	**75 %**
Fins. (T2), AC	**80 %**	**50 %**	**100 %**	**55 %**	**50 %**	**75 %**
Fins. (T2), ABC	**83 %**	**13 %**	**100 %**	**54 %**	**50 %**	75 %
Hah. (T2), AC	**80 %**	**50 %**	**100 %**	**55 %**	**50 %**	**75 %**
Hah. (T2), ABC	**83 %**	**11 %**	**100 %**	**52 %**	**50 %**	75 %
Seul. (T3), AC	**100 %**	**100 %**	**100 %**	**55 %**	**50 %**	**75 %**
Seul. (T3), ABC	**100 %**	**18 %**	**100 %**	**52 %**	**50 %**	75 %
Perusm.h. (T4), A	**88 %**	**100 %**	**100 %**	**75 %**	**100 %**	**60 %**
Perusm.h. (T4), AB	**90 %**	**27 %**	**100 %**	**64 %**	**100 %**	60 %
Perusm.h. (T4), AC	**80 %**	**50 %**	**100 %**	**55 %**	**50 %**	**75 %**
Perusm.h. (T4), ABC	**83 %**	**13 %**	**100 %**	**54 %**	**50 %**	75 %
Osit. p.m.h. (T5), A	**88 %**	**100 %**	**100 %**	**75 %**	**100 %**	**60 %**
Osit. p.m.h. (T5), AB	**90 %**	**27 %**	**100 %**	**64 %**	**100 %**	60 %
Osit. p.m.h. (T5), AC	**80 %**	**43 %**	**100 %**	**55 %**	**50 %**	**75 %**
Osit. p.m.h. (T5), ABC	**83 %**	**11 %**	**100 %**	**52 %**	**50 %**	75 %
Perint. (T1), ABCabc					**50 %**	**38 %**
Fins. (T2), ACac					**50 %**	**38 %**
Fins. (T2), ABCabc					**50 %**	**38 %**
Hah. (T2), ACac					**50 %**	**38 %**
Hah. (T2), ABCabc					**50 %**	**38 %**
Seul. (T3), ACac					**50 %**	**60 %**
Seul. (T3), ABCabc					**50 %**	**46 %**
Per.m.h. (T4), Aa					**100 %**	**60 %**
Per.m.h. (T4), ABab					**100 %**	**38 %**
Per.m.h. (T4), ACac					**50 %**	**38 %**
Per.m.h. (T4), ABCabc					**50 %**	**38 %**
Os. p.m.h. (T5), Aa					**100 %**	**60 %**
Os.p.m.h. (T5), ABab					**100 %**	**38 %**
Os.p.m.h. (T5), ACac	**67 %**				**50 %**	**50 %**
Os.p.m.h. (T5), ABCabc	**71 %**				**50 %**	**38 %**

Tutkimusympäristö	**J**		**Y**	**Y**	**Y**	**Y**
ja kyselytyyppi	**13**	**14**	**15**	**16**	**17**	**19**
Perint. (T1), ABC	**63 %**	**90 %**	**33 %**	**100 %**	**79 %**	**86 %**
Fins. (T2), AC	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**86 %**
Fins. (T2), ABC	**63 %**	90 %	33 %	100 %	79 %	86 %
Hah. (T2), AC	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**86 %**
Hah. (T2), ABC	**63 %**	90 %	33 %	100 %	79 %	86 %
Seul. (T3), AC	**80 %**	**89 %**	**33 %**	**100 %**	**79 %**	**83 %**
Seul. (T3), ABC	**63 %**	**89 %**	33 %	100 %	79 %	83 %
Perusm.h. (T4), A	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**83 %**
Perusm.h. (T4), AB	**64 %**	90 %	33 %	100 %	79 %	83 %
Perusm.h. (T4), AC	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**86 %**
Perusm.h. (T4), ABC	**63 %**	90 %	33 %	100 %	79 %	86 %
Osit. p.m.h. (T5), A	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**83 %**
Osit. p.m.h. (T5), AB	**64 %**	90 %	33 %	100 %	79 %	83 %
Osit. p.m.h. (T5), AC	**80 %**	**90 %**	**33 %**	**100 %**	**79 %**	**83 %**
Osit. p.m.h. (T5), ABC	**63 %**	90 %	33 %	100 %	79 %	83 %
Perint. (T1), ABCabc	**51 %**	**15 %**	**50 %**	**100 %**	**75 %**	**82 %**
Fins. (T2), ACac	**67 %**	**83 %**	**50 %**	**100 %**	**78 %**	**82 %**
Fins. (T2), ABCabc	**51 %**	**63 %**	**50 %**	**100 %**	**78 %**	**82 %**
Hah. (T2), ACac	**67 %**	**83 %**	**50 %**	**100 %**	**78 %**	**82 %**
Hah. (T2), ABCabc	**51 %**	**65 %**	**50 %**	**100 %**	**78 %**	**82 %**
Seul. (T3), ACac	**67 %**	**89 %**	**50 %**	**100 %**	**78 %**	**75 %**
Seul. (T3), ABCabc	**56 %**	**82 %**	**50 %**	**100 %**	**78 %**	**80 %**
Per.m.h. (T4), Aa	**80 %**	**90 %**	**56 %**	**100 %**	**78 %**	**83 %**
Per.m.h. (T4), ABab	**59 %**	**80 %**	**56 %**	**100 %**	**78 %**	**86 %**
Per.m.h. (T4), ACac	**67 %**	**86 %**	**50 %**	**100 %**	**78 %**	**82 %**
Per.m.h. (T4), ABCabc	**51 %**	**63 %**	**50 %**	**100 %**	**78 %**	**82 %**
Os. p.m.h. (T5), Aa	**80 %**	**90 %**	**56 %**	**100 %**	**78 %**	**83 %**
Os.p.m.h. (T5), ABab	**59 %**	**80 %**	**56 %**	**100 %**	**78 %**	**86 %**
Os.p.m.h. (T5), ACac	**62 %**	**60 %**	**50 %**	**100 %**	**71 %**	**78 %**
Os.p.m.h. (T5), ABCabc	**50 %**	**50 %**	**50 %**	**100 %**	**71 %**	**82 %**

Tutkimusympäristö	**Y+J**				**Y**	
ja kyselytyyppi	**20**	**21**	**24**	**25**	**26**	**27**
Perint. (T1), ABC	**80 %**	**75 %**	**100 %**	**100 %**	**100 %**	**78 %**
Fins. (T2), AC	**50 %**	**75 %**	**100 %**	**100 %**	**100 %**	**78 %**
Fins. (T2), ABC	**80 %**	75 %	100 %	100 %	100 %	78 %
Hah. (T2), AC	**50 %**	**75 %**	**100 %**	**100 %**	**100 %**	**78 %**
Hah. (T2), ABC	**80 %**	75 %	100 %	100 %	100 %	78 %
Seul. (T3), AC	**50 %**	**75 %**	**100 %**	**100 %**	**100 %**	**88 %**
Seul. (T3), ABC	**80 %**	75 %	100 %	100 %	100 %	88 %
Perusm.h. (T4), A	**100 %**	**75 %**	**100 %**	**100 %**	**100 %**	**88 %**
Perusm.h. (T4), AB	**100 %**	75 %	100 %	100 %	100 %	88 %
Perusm.h. (T4), AC	**100 %**	75 %	**100 %**	**100 %**	**100 %**	**78 %**
Perusm.h. (T4), ABC	**100 %**	75 %	100 %	100 %	100 %	78 %
Osit. p.m.h. (T5), A	**100 %**	**75 %**	**100 %**	**100 %**	**100 %**	**88 %**
Osit. p.m.h. (T5), AB	**100 %**	75 %	100 %	100 %	100 %	88 %
Osit. p.m.h. (T5), AC	**100 %**	75 %	**100 %**	**100 %**	**100 %**	**89 %**
Osit. p.m.h. (T5), ABC	**92 %**	75 %	100 %	100 %	100 %	89 %
Perint. (T1), ABCabc	**80 %**				**100 %**	
Fins. (T2), ACac	**50 %**				**100 %**	
Fins. (T2), ABCabc	**80 %**				100 %	
Hah. (T2), ACac	**50 %**				**100 %**	
Hah. (T2), ABCabc	**80 %**				100 %	
Seul. (T3), ACac	**50 %**				**100 %**	
Seul. (T3), ABCabc	**80 %**				100 %	
Per.m.h. (T4), Aa	**100 %**				**100 %**	
Per.m.h. (T4), ABab	**100 %**				100 %	
Per.m.h. (T4), ACac	**100 %**				**100 %**	
Per.m.h. (T4), ABCabc	**100 %**				100 %	
Os. p.m.h. (T5), Aa	**100 %**				**100 %**	
Os.p.m.h. (T5), ABab	**100 %**				100 %	
Os.p.m.h. (T5), ACac	**100 %**		**100 %**		**100 %**	**80 %**
Os.p.m.h. (T5), ABCabc	**93 %**		100 %		100 %	80 %

Tutkimusympäristö ja kyselytyyppi	**28**	**29**
Perint. (T1), ABC	**67 %**	**100 %**
Fins. (T2), AC	**67 %**	**100 %**
Fins. (T2), ABC	67 %	100 %
Hah. (T2), AC	**67 %**	**100 %**
Hah. (T2), ABC	67 %	100 %
Seul. (T3), AC		**100 %**
Seul. (T3), ABC		100 %
Perusm.h. (T4), A	**100 %**	**100 %**
Perusm.h. (T4), AB	100 %	100 %
Perusm.h. (T4), AC	**67 %**	**100 %**
Perusm.h. (T4), ABC	67 %	100 %
Osit. p.m.h. (T5), A	**100 %**	**100 %**
Osit. p.m.h. (T5), AB	100 %	100 %
Osit. p.m.h. (T5), AC	**67 %**	**100 %**
Osit. p.m.h. (T5), ABC	67 %	100 %
Perint. (T1), ABCabc		**29 %**
Fins. (T2), ACac		**50 %**
Fins. (T2), ABCabc		**29 %**
Hah. (T2), ACac		**50 %**
Hah. (T2), ABCabc		**29 %**
Seul. (T3), ACac		**86 %**
Seul. (T3), ABCabc		**67 %**
Per.m.h. (T4), Aa		**86 %**
Per.m.h. (T4), ABab		**75 %**
Per.m.h. (T4), ACac		**50 %**
Per.m.h. (T4), ABCabc		**29 %**
Os. p.m.h. (T5), Aa		**86 %**
Os.p.m.h. (T5), ABab		**75 %**
Os.p.m.h. (T5), ACac	**67 %**	**86 %**
Os.p.m.h. (T5), ABCabc	67 %	**64 %**

N_p = 26, perusjoukon koko

N_y = 9, yhdyssanaosajoukon koko

N_j = 8, johdososajoukon koko

			Y	Y	J	J
Tutkimusympäristö ja kyselytyyppi	**N_p**	**Keski-arvo**	**N_y**	**Keski-arvo**	**N_j**	**Keski-arvo**
Perint. (T1), ABC	26	**76,6 %**	9	**80,0 %**	8	**64,3 %**
Fins. (T2), AC	26	77,3 %	9	76,7 %	8	66,5 %
Fins. (T2), ABC	26	**76,6 %**	9	**80,0 %**	8	**64,3 %**
Hah. (T2), AC	26	77,3 %	9	76,7 %	8	66,5 %
Hah. (T2), ABC	26	**76,5 %**	9	**80,0 %**	8	**63,9 %**
Seul. (T3), AC	25	80,7 %	9	76,4 %	8	75,2 %
Seul. (T3), ABC	25	**78,4 %**	9	**79,8 %**	8	**67,9 %**
Perusm.h. (T4), A	26	84,2 %	9	87,5 %	8	88,6 %
Perusm.h. (T4), AB	26	80,4 %	9	87,5 %	8	76,4 %
Perusm.h. (T4), AC	26	79,2 %	9	82,3 %	8	72,7 %
Perusm.h. (T4), ABC	26	**77,3 %**	9	**82,3 %**	8	**66,8 %**
Osit. p.m.h. (T5), A	26	84,2 %	9	87,5 %	8	88,6 %
Osit. p.m.h. (T5), AB	26	80,4 %	9	87,5 %	8	76,4 %
Osit. p.m.h. (T5), AC	26	79,2 %	9	82,0 %	8	71,8 %
Osit. p.m.h. (T5), ABC	26	**77,3 %**	9	**81,1 %**	8	**65,4 %**
Perint. (T1), ABCabc	14	**66,3 %**	9	**81,1 %**		
Fins. (T2), ACac	14	71,8 %	9	78,1 %		
Fins. (T2), ABCabc	14	**69,9 %**	9	**81,4 %**		
Hah. (T2), ACac	14	71,8 %	9	78,1 %		
Hah. (T2), ABCabc	14	**70,0 %**	9	**81,4 %**		
Seul. (T3), ACac	14	76,3 %	9	77,3 %		
Seul. (T3), ABCabc	14	**75,2 %**	9	**81,2 %**		
Per.m.h. (T4), Aa	14	85,4 %	9	89,9 %		
Per.m.h. (T4), ABab	14	80,9 %	9	90,2 %		
Per.m.h. (T4), ACac	14	75,4 %	9	83,7 %		
Per.m.h. (T4), ABCabc	14	**71,0 %**	9	**83,7 %**		
Os. p.m.h. (T5), Aa	14	85,4 %	9	89,9 %		
Os.p.m.h. (T5), ABab	14	80,9 %	9	90,2 %		
Os.p.m.h. (T5), ACac	14	76,0 %	9	82,5 %		
Os.p.m.h. (T5), ABCabc	14	**71,7 %**	9	**82,1 %**		

ESIMERKKI MERKITSEVYYSTESTILASKELMASTA

T4, perusjoukko, saanti, JA-operaattori

N = **26** (Hakupyyntöjen eli rivien määrä)

k = **5** (Vertailtavien kyselytyyppien määrä)

Kyselytyypin järjestysnumero

Haku-pyynnön numero	**T1/ABC**	**T4/A**	**T4/AB**	**T4/AC**	**T4/ABC**	Järjestys-lukujen neliöiden summa
1	3	3	3	3	3	45,0
2	4	1,5	1,5	4	4	52,5
3	4	1,5	1,5	4	4	52,5
4	3	3	3	3	3	45,0
5	4	1,5	1,5	4	4	52,5
6	4,5	1	3	2	4,5	54,5
7	4,5	1	2,5	2,5	4,5	54,0
8	4,5	1,5	3	1,5	4,5	54,0
9	4	1,5	4	1,5	4	52,5
10	4,5	1	2	3	4,5	54,5
11	3	3	3	3	3	45,0
12	4	1,5	1,5	4	4	52,5
13	4,5	1,5	3	1,5	4,5	54,0
14	5	2,5	2,5	2,5	2,5	50,0
15	3	3	3	3	3	45,0
16	3	3	3	3	3	45,0
17	3	3	3	3	3	45,0
19	4	1,5	1,5	4	4	52,5
20	3,5	1	3,5	2	5	54,5
21	3	3	3	3	3	45,0
24	3	3	3	3	3	45,0
25	3	3	3	3	3	45,0
26	3	3	3	3	3	45,0
27	4	1,5	1,5	4	4	52,5
28	3	3	3	3	3	45,0
29	3	3	3	3	3	45,0
$\boldsymbol{R_j}$	95	56	68,5	76,5	94	**1283,0** = **A**
$\boldsymbol{R_j^2}$	9025	3136	4692,25	5852,25	8836	**31541,5** = **S$\boldsymbol{R_j^2}$**

järjestys-lukujen summien neliöiden summa

B = $\frac{1}{N}\sum_{j=1}^{k} R_j^2$ siis: **B = 1 213,1**

T = $\frac{(N-1)\left[B - Nk(k+1)^2/4\right]}{A-B}$ **T = 15,4**

T:n arvon ja vapausasteiden

k_1 = **4** k_2 = **100**

mukaan voidaan todeta, että joidenkin ryhmien väliset erot
ovat merkitseviä merkitsevyystasolla 0.01.
(Conover 1980, liitteen A26-taulukko)
Vertailtavien ryhmien järjestyslukujen summien erotuksen on oltava
suurempi kuin vertailuluku, joka saadaan kaavalla:

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

t :n arvo kun merkitsevyystaso = 0.01 eli α = 0.99 ja vapausaste (N-1)(k-1) = 100:

$t_{.99}$ = **2,36867** (Conover 1980, interpoloitu liitteen A25-taulukosta)

=> Vertailuluku = **14,3**

t :n arvo kun merkitsevyystaso = 0.05 eli α = 0.95 ja vapausaste = 100:

$t_{.95}$ = **1,66233**

=> Vertailuluku = **10,0**

t :n arvo kun merkitsevyystaso = 0.1 eli α = 0.90 ja vapausaste = 100:

$t_{.90}$ = **1,29133**

=> Vertailuluku = **7,8**

Itseisarvojen erotukset:

	T4/A	**T4/AB**	**T4/AC**	**T4/ABC**
T1/ABC	**39**	**26,5**	**18,5**	1
T4/A		***12,5***	**20,5**	**38**
T4/AB			8	**25,5**
T4/AC				**17,5**

Merkitsevyystaso 0.01:
T4/A muista paitsi johdoskyselystä eli:
T4/A - T1/ABC, T4/A - T4/AC, T4/A - T4/ABC
T1/ABC muista paitsi vastaavasta T4-yhdistelmäkyselystä eli:
T4/AB - T1/ABC, T4/AC - T1/ABC
T4/ABC muista T4-ympäristön kyselyistä eli:
T4/AB - T4/ABC, T4/AC - T4/ABC

Merkitsevyystaso 0.05:
T4/A - T4/AB

Merkitsevyystaso 0.1:
T4/AB - T4/AC

MERKITSEVYYSTESTIEN TULOSTEN YHTEENVETO

Liitteessä esitetään Friedmanin merkitsevyystestien tulokset eri tutkimusjoukoissa saannin ja tarkkuuden osalta:
tutkimusympäristöt ja kyselytyypit, joiden saanti/tarkkuusarvojen välinen ero oli tilastollisesti merkitsevä, ja millä tilastollisella merkitsevyystasolla. Tässä siis on listattu vain tilastollisesti merkitsevät erot.

Luvuissa 1 – 3 muiden tutkimusympäristöjen vertailukohtana oli **T1** eli perinteinen hakutapa taivutusmuotohakemistosta. T2-ympäristön tulokset on saatu **Finstems**-ohjelmaa käyttäen. Kunkin tutkimusympäristön kyselytyyppejä on vertailtu myös keskenään.
Luvussa 4 vertaillaan **kaikkien** ympäristöjen saantia ja tarkkuutta kyselytyypillä, jota käytettiin kaikissa tutkimusympäristöissä, eli (osien) yhdistelmäkyselyllä.

Taulukossa on seuraavat sarakkeet:
- TY: tutkimusympäristön lyhenne
- JA/Virke: operaattori, jolla hakusanat kyselyssä kytkettiin toisiinsa
- i ja j: vertailtavat kyselytyypit
- α: merkitsevyystaso
- $\overline{x_i} - \overline{x_j}$: vertailtujen kyselytyyppien saanti/tarkkuusarvojen keskiarvojen välinen ero
- SJ: Sparck Jonesin käytännön merkitsevyys: 5 – 10 prosenttiyksikön ero on huomattava (H), yli 10 prosenttiyksikön ero olennainen (O)

1 PERUSJOUKKO

Vertailtujen kyselyjen määrä N = 26, paitsi T3-ympäristön vertailuissa N = 25.

TY	JA/ Virke	SAANTI i		j	α	$x_i - x_j$	SJ	TARKKUUS i		j	α	$x_i - x_j$	SJ
T2	JA	T2/AC	<	T1/ABC	**0.01**	8,1	H	T2/AC	>	T1/ABC	**0.025**	3,1	
		T2/AC	<	T2/ABC	**0.01**	7,9	H	T2/AC	>	T2/ABC	**0.025**	2,4	
	Virke	T2/AC	<	T1/ABC	**0.01**	5,9	H	(ei merkit-					
		T2/AC	<	T2/ABC	**0.01**	5,8	H	seviä eroja)					
T3	JA	T3/AC	<	T1/ABC	**0.01**	10,3	O	T3/AC	>	T1/ABC	**0.05**	4,4	
		T3/AC	<	T3/ABC	**0.05**	6,5	H	T3/AC	>	T3/ABC	**0.05**	3,0	
		T3/ABC	<	T1/ABC	**0.05**	3,8							
	Virke	T3/AC	<	T1/ABC	**0.01**	7,8	H	(ei merkit-					
		T3/AC	<	T3/ABC	**0.01**	5,1	H	seviä eroja)					
		T3/ABC	<	T1/ABC	**0.05**	2,7							
T4	JA	T4/A	<	T1/ABC	**0.01**	17,3	O	T4/A	>	T1/ABC	**0.01**	7,8	H
		T4/A	<	T4/AC	**0.01**	9,4	H	T4/A	>	T4/ABC	**0.01**	5,5	H
		T4/A	<	T4/ABC	**0.01**	17,3	O	T4/AC	>	T1/ABC	**0.01**	5,0	H
		T4/AB	<	T1/ABC	**0.01**	10,3	O	T4/AB	>	T1/ABC	**0.05**	4,4	
		T4/AB	<	T4/ABC	**0.01**	10,3	O	T4/A	>	T4/AB	0.1	3,4	
		T4/AC	<	T1/ABC	**0.01**	7,9	H	T4/AC	>	T4/ABC	0.1	2,7	
		T4/AC	<	T4/ABC	**0.01**	7,9	H	T4/ABC	>	T1/ABC	0.1	2,3	
		T4/A	<	T4/AB	**0.05**	7,0	H						
		T4/AB	<	T4/AC	0.1	2,4							

TY	JA/ Virke	SAANTI					TARKKUUS				
		i	j	α	$x_i - x_j$	SJ	i	j	α	$x_i - x_j$	SJ
T4	Virke	T4/A <	T1/ABC	**0.01**	13,8	O	(ei merkit-				
		T4/A <	T4/AB	**0.01**	5,7	H	seviä eroja)				
		T4/A <	T4/AC	**0.01**	8,1	H					
		T4/A <	T4/ABC	**0.01**	13,9	O					
		T4/AB <	T1/ABC	**0.01**	8,1	H					
		T4/AB <	T4/ABC	**0.01**	8,2	H					
		T4/AC <	T1/ABC	**0.01**	5,7	H					
		T4/AC <	T4/ABC	**0.01**	5,8	H					
T5	JA	T5/A <	T1/ABC	**0.01**	17,3	O	T5/A >	T1/ABC	**0.01**	7,8	H
		T5/A <	T5/AC	**0.01**	11,0	O	T5/AB >	T1/ABC	**0.01**	4,4	
		T5/A <	T5/ABC	**0.01**	20,4	O	T5/AC >	T1/ABC	**0.01**	5,2	H
		T5/AB <	T1/ABC	**0.01**	10,3	O	T5/A >	T5/ABC	**0.05**	4,8	
		T5/AB <	T5/ABC	**0.01**	13,4	O	T5/AC >	T5/ABC	0.1	2,2	
		T5/AC <	T5/ABC	**0.01**	9,4	H	T5/ABC >	T1/ABC	0.1	3,0	
		T5/A <	T5/AB	**0.05**	7,0	H					
		T5/AB <	T5/AC	**0.05**	4,0						
		T5/AC <	T1/ABC	**0.05**	6,3	H					
	Virke	T5/A <	T1/ABC	**0.01**	13,8	O	(ei merkit-				
		T5/A <	T5/AB	**0.01**	5,7	H	seviä eroja)				
		T5/A <	T5/AC	**0.01**	8,3	H					
		T5/A <	T5/ABC	**0.01**	14,7	O					
		T5/AB <	T1/ABC	**0.01**	8,1	H					
		T5/AB <	T5/ABC	**0.01**	9,0	H					
		T5/AC <	T1/ABC	**0.01**	5,5	H					
		T5/AC <	T5/ABC	**0.01**	6,4	H					

2 JOHDOSOSAJOUKKO

Kyselyjen määrä N = 8.

TY	JA/ Virke	SAANTI i		j	α	$x_i - x_j$	SJ	TARKKUUS i		j	α	$x_i - x_j$	SJ
T2	JA	T2/AC	<	T1/ABC	**0.01**	25,6	O	T2/AC	>	T1/ABC	**0.025**	7,8	H
		T2/AC	<	T2/ABC	**0.01**	25,6	O	T2/AC	>	T2/ABC	**0.025**	7,8	H
	Virke	T2/AC	<	T1/ABC	**0.01**	18,9	O	(ei merkit-					
		T2/AC	<	T2/ABC	**0.01**	18,9	O	seviä eroja)					
T3	JA	T3/AC	<	T1/ABC	**0.01**	29,3	O	T3/AC	>	T1/ABC	**0.05**	10,8	O
		T3/AC	<	T3/ABC	**0.01**	20,3	O	T3/AC	>	T3/ABC	**0.05**	9,4	H
		T3/ABC	<	T1/ABC	**0.05**	9	H						
	Virke	T3/AC	<	T1/ABC	**0.01**	22,5	O	(ei merkit-					
		T3/AC	<	T3/ABC	**0.01**	16,2	O	seviä eroja)					
		T3/ABC	<	T1/ABC	**0.05**	6,3	H						
T4	JA	T4/A	<	T1/ABC	**0.01**	30,8	O	T4/A	>	T1/ABC	**0.01**	20,4	O
		T4/A	<	T4/AB	**0.01**	22,6	O	T4/A	>	T4/ABC	**0.01**	19,1	O
		T4/A	<	T4/ABC	**0.01**	31,5	O	T4/AC	>	T1/ABC	**0.01**	10,3	O
		T4/AB	<	T1/ABC	**0.01**	8,2	H	T4/A	>	T4/AB	**0.05**	10,8	O
		T4/AB	>	T4/AC	**0.01**	16,8	O	T4/A	>	T4/AC	**0.05**	10,1	O
		T4/AB	<	T4/ABC	**0.01**	8,9	H	T4/AB	>	T1/ABC	**0.05**	9,6	H
		T4/AC	<	T1/ABC	**0.01**	25,0	O	T4/AB	>	T4/ABC	**0.05**	8,3	H
		T4/AC	<	T4/ABC	**0.01**	25,7	O	T4/AC	>	T4/ABC	**0.05**	9,0	H
		T4/A	<	T4/AC	**0.05**	5,8	H						

TY	JA/ Virke	SAANTI					TARKKUUS				
		i	j	α	$x_i - x_j$	SJ	i	j	α	$x_i - x_j$	SJ
T4	Virke	T4/A <	T1/ABC	**0.01**	21,9	O	T4/A >	T1/ABC	**0.05**	24,3	O
		T4/A <	T4/AB	**0.01**	18,5	O	T4/AB >	T1/ABC	**0.05**	12,1	O
		T4/A <	T4/ABC	**0.01**	22,6	O	T4/A >	T4/AC	**0.05**	15,9	O
		T4/AB >	T4/AC	**0.01**	14,8	O	T4/A >	T4/ABC	**0.05**	21,8	O
		T4/AB <	T4/ABC	**0.01**	4,1		T4/AB >	T4/ABC	0.1	9,6	H
		T4/AC <	T1/ABC	**0.01**	18,2	O					
		T4/AC <	T4/ABC	**0.01**	18,9	O					
		T4/AB <	T1/ABC	**0.05**	3,4						
T5	JA	T5/A <	T1/ABC	**0.01**	30,8	O	T5/A >	T1/ABC	**0.01**	20,4	O
		T5/A <	T5/AB	**0.01**	22,6	O	T5/A >	T5/ABC	**0.01**	19,5	O
		T5/A <	T5/ABC	**0.01**	36,9	O	T5/AC >	T1/ABC	**0.01**	8,0	H
		T5/AB <	T1/ABC	**0.01**	8,2	H	T5/A >	T5/AB	**0.05**	10,8	O
		T5/AB <	T5/ABC	**0.01**	14,3	O	T5/A >	T5/AC	**0.05**	12,4	O
		T5/AC <	T1/ABC	**0.01**	24,5	O	T5/AB >	T1/ABC	**0.05**	9,6	H
		T5/AC <	T5/ABC	**0.01**	30,6	O	T5/AB >	T5/ABC	**0.05**	8,7	H
		T5/A <	T5/AC	**0.05**	6,3	H	T5/AC >	T5/ABC	**0.05**	7,1	H
		T5/AB >	T5/AC	**0.05**	16,3	O					
	Virke	T5/A <	T1/ABC	**0.01**	21,9	O	T5/A >	T1/ABC	**0.05**	24,3	O
		T5/A <	T5/AB	**0.01**	18,5	O	T5/A >	T5/AC	**0.05**	16,8	O
		T5/A <	T5/ABC	**0.01**	24,7	O	T5/A >	T5/ABC	**0.05**	23,2	O
		T5/AB >	T5/AC	**0.01**	14,8	O	T5/AB >	T1/ABC	**0.05**	12,1	O
		T5/AC <	T1/ABC	**0.01**	18,2	O	T5/AB >	T5/ABC	**0.05**	11,0	O
		T5/AC <	T5/ABC	**0.01**	21,0	O					
		T5/AB <	T1/ABC	**0.05**	3,4						
		T5/AB <	T5/ABC	**0.05**	6,2	H					

3 YHDYSSANAOSAJOUKKO

Kyselyjen määrä N = 9.

TY	JA/ Virke	SAANTI i		j	α	$x_i - x_j$	SJ	TARKKUUS i		j	α	$x_i - x_j$	SJ
T2	JA	T1/ABC	<	T1/ABCabc	**0.01**	32,4	O	T1/ABC	>	T1/ABCabc	**0.01**	34,0	O
		T1/ABC	<	T2/ACac	**0.01**	27,4	O	T1/ABC	>	T2/ABCabc	**0.01**	33,1	O
		T1/ABC	<	T2/ABCabc	**0.01**	32,4	O	T2/AC	>	T1/ABCabc	**0.01**	36,3	O
		T2/AC	<	T1/ABCabc	**0.01**	34,1	O	T2/AC	>	T2/ABCabc	**0.01**	35,4	O
		T2/AC	<	T2/ACac	**0.01**	29,1	O	T2/ABC	>	T1/ABCabc	**0.01**	36,1	O
		T2/AC	<	T2/ABCabc	**0.01**	34,1	O	T2/ABC	>	T2/ABCabc	**0.01**	35,2	O
		T2/ABC	<	T1/ABCabc	**0.01**	32,4	O	T2/AC	>	T2/ACac	**0.05**	31,3	O
		T2/ABC	<	T2/ACac	**0.01**	27,4	O	T2/ABC	>	T2/ACac	**0.05**	31,1	O
		T2/ABC	<	T2/ABCabc	**0.01**	32,4	O	T2/ACac	>	T1/ABCabc	**0.05**	5,0	H
		T2/ACac	<	T1/ABCabc	0.1	5,0	H	T2/ACac	>	T2/ABCabc	0.1	4,1	
		T2/ACac	<	T2/ABCabc	0.1	5,0	H						
	Virke	T1/ABC	<	T1/ABCabc	**0.01**	14,0	O	(ei merkit-					
		T1/ABC	<	T2/ACac	**0.01**	12,2	O	seviä eroja)					
		T1/ABC	<	T2/ABCabc	**0.01**	14,0	O						
		T2/AC	<	T1/ABCabc	**0.01**	15,7	O						
		T2/AC	<	T2/ACac	**0.01**	13,9	O						
		T2/AC	<	T2/ABCabc	**0.01**	15,7	O						
		T2/ABC	<	T1/ABCabc	**0.01**	14,0	O						
		T2/ABC	<	T2/ACac	**0.01**	13,9	O						
		T2/ABC	<	T2/ABCabc	**0.01**	14,0	O						

TY	JA/ Virke	SAANTI i		j	α	$x_i - x_j$	SJ	TARKKUUS i		j	α	$x_i - x_j$	SJ
T3	JA	T1/ABC	<	T1/ABCabc	**0.01**	32,4	O	T1/ABC	>	T1/ABCabc	**0.01**	34,0	O
		T1/ABC	<	T3/ACac	**0.01**	21,1	O	T3/AC	>	T1/ABCabc	**0.01**	35,8	O
		T1/ABC	<	T3/ABCabc	**0.01**	27,8	O	T3/ABC	>	T1/ABCabc	**0.01**	35,6	O
		T3/AC	<	T1/ABCabc	**0.01**	35,3	O	T3/ACac	>	T1/ABCabc	**0.01**	11,7	O
		T3/AC	<	T3/ACac	**0.01**	24,0	O	T3/ABCabc	>	T1/ABCabc	**0.01**	8,3	H
		T3/AC	<	T3/ABCabc	**0.01**	30,7	O	T3/AC	>	T3/ABCabc	0.1	27,5	O
		T3/ABC	<	T1/ABCabc	**0.01**	33,6	O						
		T3/ABC	<	T3/ACac	**0.01**	22,3	O						
		T3/ABC	<	T3/ABCabc	**0.01**	29,0	O						
		T3/ACac	<	T1/ABCabc	**0.01**	11,3	O						
		T3/ACac	<	T3/ABCabc	**0.05**	6,7	H						
	Virke	T1/ABC	<	T1/ABCabc	**0.01**	14,0	O	(ei merkit-					
		T1/ABC	<	T3/ABCabc	**0.01**	13,4	O	seviä eroja)					
		T3/AC	<	T1/ABCabc	**0.01**	16,3	O						
		T3/AC	<	T3/ACac	**0.01**	12,8	O						
		T3/AC	<	T3/ABCabc	**0.01**	15,7	O						
		T3/ABC	<	T1/ABCabc	**0.01**	14,5	O						
		T3/ABC	<	T3/ACac	**0.01**	11,0	O						
		T3/ABC	<	T3/ABCabc	**0.01**	13,9	O						
		T1/ABC	<	T3/ACac	**0.05**	10,5	O						

TY	JA/ Virke	SAANTI						TARKKUUS					
		i		j	α	$x_i - x_j$	SJ	i		j	α	$x_i - x_j$	SJ
T4	JA	T1/ABC	<	T1/ABCabc	**0.01**	32,4	O	T1/ABC	>	T1/ABCabc	**0.01**	34,0	O
		T1/ABC	<	T4/ACac	**0.01**	28,0	O	T1/ABC	>	T4/ABCabc	**0.01**	31,5	O
		T1/ABC	<	T4/ABCabc	**0.01**	32,9	O	T4/A	>	T1/ABCabc	**0.01**	44,0	O
		T4/A	<	T1/ABCabc	**0.01**	39,3	O	T4/A	>	T4/ACac	**0.01**	36,6	O
		T4/A	<	T4/Aa	**0.01**	18,6	O	T4/A	>	T4/ABCabc	**0.01**	41,5	O
		T4/A	<	T4/ABab	**0.01**	24,2	O	T4/AB	>	T1/ABCabc	**0.01**	42,6	O
		T4/A	<	T4/ACac	**0.01**	34,9	O	T4/AB	>	T4/ACac	**0.01**	35,2	O
		T4/A	<	T4/ABCabc	**0.01**	39,8	O	T4/AB	>	T4/ABCabc	**0.01**	40,1	O
		T4/AB	<	T1/ABCabc	**0.01**	36,9	O	T4/AC	>	T1/ABCabc	**0.01**	38,6	O
		T4/AB	<	T4/ABab	**0.01**	21,8	O	T4/AC	>	T4/ACac	**0.01**	31,2	O
		T4/AB	<	T4/ACac	**0.01**	32,5	O	T4/AC	>	T4/ABCabc	**0.01**	36,1	O
		T4/AB	<	T4/ABCabc	**0.01**	37,4	O	T4/ABC	>	T1/ABCabc	**0.01**	37,3	O
		T4/AC	<	T1/ABCabc	**0.01**	33,6	O	T4/ABC	>	T4/ACac	**0.01**	29,9	O
		T4/AC	<	T4/ABab	**0.01**	18,5	O	T4/ABC	>	T4/ABCabc	**0.01**	34,8	O
		T4/AC	<	T4/ACac	**0.01**	29,2	O	T4/Aa	>	T1/ABCabc	**0.01**	23,6	O
		T4/AC	<	T4/ABCabc	**0.01**	34,1	O	T4/ABab	>	T1/ABCabc	**0.01**	14,0	O
		T4/ABC	<	T1/ABCabc	**0.01**	31,8	O	T4/Aa	>	T4/ABCabc	**0.01**	21,1	O
		T4/ABC	<	T4/ACac	**0.01**	27,4	O	T4/A	>	T4/ABab	**0.05**	30,0	O
		T4/ABC	<	T4/ABCabc	**0.01**	32,3	O	T4/AB	>	T4/ABab	**0.05**	28,6	O
		T4/Aa	<	T1/ABCabc	**0.01**	20,7	O	T4/AC	>	T4/ABab	**0.05**	24,6	O
		T4/ABab	<	T1/ABCabc	**0.01**	15,1	O	T4/ABC	>	T4/ABab	**0.05**	23,3	O
		T4/Aa	<	T4/ACac	**0.01**	16,3	O	T4/ACac	>	T1/ABCabc	**0.05**	7,4	H
		T4/Aa	<	T4/ABCabc	**0.01**	21,2	O	T4/ABab	>	T4/ABCabc	**0.05**	11,5	O
		T4/ABab	<	T4/ABCabc	**0.01**	15,6	O	T4/AB	>	T1/ABC	0.1	8,6	H
		T4/A	<	T1/ABC	**0.05**	6,9	H	T4/AC	>	T1/ABC	0.1	4,6	
		T1/ABC	<	T4/ABab	**0.05**	17,3	O	T1/ABC	>	T4/ACac	0.1	26,6	O
		T4/A	<	T4/ABC	**0.05**	6,9	H	T4/AB	>	T4/Aa	0.1	19,0	O
		T4/AB	<	T4/Aa	**0.05**	16,2	O	T4/AC	>	T4/Aa	0.1	15,0	O

		T4/ABC < T4/ABab	**0.05**	16,7	O	T4/Aa > T4/ACac	0.1	16,2	O
		T4/ABab < T4/ACac	**0.05**	10,7	O				
		T4/AB < T4/ABC	0.1	5,1	H				
		T4/Aa < T4/ABab	0.1	5,6	H				
		T4/ACac < T4/ABCabc	0.1	4,9					
T4	Virke	T1/ABC < T1/ABCabc	**0.01**	14,0	O	(ei merkit-			
		T1/ABC < T4/ACac	**0.01**	12,8	O	seviä eroja)			
		T1/ABC < T4/ABCabc	**0.01**	14,6	O				
		T4/A < T1/ABCabc	**0.01**	15,7	O				
		T4/A < T4/ABab	**0.01**	11,4	O				
		T4/A < T4/ACac	**0.01**	14,5	O				
		T4/A < T4/ABCabc	**0.01**	16,3	O				
		T4/AB < T1/ABCabc	**0.01**	14,0	O				
		T4/AB < T4/ACac	**0.01**	12,8	O				
		T4/AB < T4/ABCabc	**0.01**	14,6	O				
		T4/AC < T1/ABCabc	**0.01**	15,1	O				
		T4/AC < T4/ABab	**0.01**	10,8	O				
		T4/AC < T4/ACac	**0.01**	13,9	O				
		T4/AC < T4/ABCabc	**0.01**	15,7	O				
		T4/ABC < T1/ABCabc	**0.01**	13,4	O				
		T4/ABC < T4/ABCabc	**0.01**	14,0	O				
		T4/Aa < T4/ABCabc	**0.01**	7,3	H				
		T1/ABC < T4/ABab	**0.05**	9,7	H				
		T4/A < T4/Aa	**0.05**	9,0	H				
		T4/AB < T4/ABab	**0.05**	9,7	H				
		T4/ABC < T4/ABab	**0.05**	9,1	H				
		T4/ABC < T4/ACac	**0.05**	12,2	O				
		T4/Aa < T1/ABCabc	**0.05**	6,7	H				
		T4/Aa < T4/ACac	**0.05**	5,5	H				
		T4/ABab < T4/ABCabc	0.1	4,9					

TY	JA/ Virke	SAANTI					TARKKUUS				
		i	j	α	$x_i - x_j$	SJ	i	j	α	$x_i - x_j$	SJ
T5	JA	T1/ABC <	T1/ABCabc	**0.01**	32,4	O	T1/ABC >	T1/ABCabc	**0.01**	34,0	O
		T1/ABC <	T5/ACac	**0.01**	28,7	O	T1/ABC >	T5/ABCabc	**0.01**	29,5	O
		T1/ABC <	T5/ABCabc	**0.01**	42,7	O	T5/A >	T1/ABCabc	**0.01**	44,0	O
		T5/A <	T1/ABCabc	**0.01**	39,3	O	T5/A >	T5/ACac	**0.01**	35,0	O
		T5/A <	T5/Aa	**0.01**	18,6	O	T5/A >	T5/ABCabc	**0.01**	39,5	O
		T5/A <	T5/ABab	**0.01**	24,2	O	T5/AB >	T1/ABCabc	**0.01**	42,6	O
		T5/A <	T5/ACac	**0.01**	35,6	O	T5/AB >	T5/ACac	**0.01**	33,6	O
		T5/A <	T5/ABCabc	**0.01**	31,0	O	T5/AB >	T5/ABCabc	**0.01**	38,1	O
		T5/AB <	T1/ABCabc	**0.01**	36,9	O	T5/AC >	T1/ABCabc	**0.01**	38,0	O
		T5/AB <	T5/ABab	**0.01**	21,8	O	T5/AC >	T5/ACac	**0.01**	29,0	O
		T5/AB <	T5/ACac	**0.01**	33,2	O	T5/AC >	T5/ABCabc	**0.01**	33,5	O
		T5/AB <	T5/ABCabc	**0.01**	47,2	O	T5/ABC >	T1/ABCabc	**0.01**	36,4	O
		T5/AC <	T1/ABCabc	**0.01**	34,7	O	T5/ABC >	T5/ABCabc	**0.01**	31,9	O
		T5/AC <	T5/ABab	**0.01**	19,6	O	T5/Aa >	T1/ABCabc	**0.01**	23,6	O
		T5/AC <	T5/ACac	**0.01**	31,0	O	T5/ABab >	T1/ABCabc	**0.01**	14,0	O
		T5/AC <	T5/ABCabc	**0.01**	45,0	O	T5/Aa >	T5/ABCabc	**0.01**	19,1	O
		T5/ABC <	T1/ABCabc	**0.01**	27,7	O	T5/A >	T1/ABC	**0.05**	10,0	O
		T5/ABC <	T5/ACac	**0.01**	24,0	O	T5/A >	T5/ABab	**0.05**	30,0	O
		T5/ABC <	T5/ABCabc	**0.01**	38,0	O	T5/AB >	T5/ABab	**0.05**	28,6	O
		T5/Aa <	T1/ABCabc	**0.01**	20,7	O	T5/ABC >	T5/ACac	**0.05**	27,4	O
		T5/ABab <	T1/ABCabc	**0.01**	15,1	O	T5/Aa >	T5/ACac	**0.05**	14,6	O
		T5/Aa <	T5/ACac	**0.01**	17,0	O	T5/ABab >	T5/ABCabc	**0.05**	9,5	H
		T5/Aa <	T5/ABCabc	**0.01**	31,0	O	T5/AB >	T1/ABC	0.1	8,6	H
		T5/ABab <	T5/ACac	**0.01**	11,4	O	T1/ABC >	T5/ACac	0.1	25,0	O
		T5/ABab <	T5/ABCabc	**0.01**	25,4	O	T5/AC >	T5/ABab	0.1	24,0	O
		T5/A <	T1/ABC	**0.05**	6,9	H	T5/ACac <	T1/ABCabc	0.1	9.0	H
		T1/ABC <	T5/ABab	**0.05**	17,3	O					
		T5/A <	T5/ABC	**0.05**	11,6	O					

		T5/AB	<	T5/Aa	**0.05**	16,2	O				
		T5/AC	<	T5/Aa	**0.05**	14,0	O				
		T5/ABC	<	T5/ABab	**0.05**	12,6	O				
		T5/AB	<	T5/ABC	0.1	9,2	H				
		T5/ACac	<	T5/ABCabc	0.1	14,0	O				
	Virke	T1/ABC	<	T1/ABCabc	**0.01**	14,0	O	(ei merkit-			
		T1/ABC	<	T5/ACac	**0.01**	15,3	O	seviä eroja)			
		T1/ABC	<	T5/ABCabc	**0.01**	22,9	O				
		T5/A	<	T1/ABCabc	**0.01**	15,7	O				
		T5/A	<	T5/ABab	**0.01**	11,4	O				
		T5/A	<	T5/ACac	**0.01**	17,0	O				
		T5/A	<	T5/ABCabc	**0.01**	24,6	O				
		T5/AB	<	T1/ABCabc	**0.01**	14,0	O				
		T5/AB	<	T5/ACac	**0.01**	15,3	O				
		T5/AB	<	T5/ABCabc	**0.01**	22,9	O				
		T5/AC	<	T1/ABCabc	**0.01**	15,7	O				
		T5/AC	<	T5/ABab	**0.01**	11,4	O				
		T5/AC	<	T5/ACac	**0.01**	17,0	O				
		T5/AC	<	T5/ABCabc	**0.01**	24,6	O				
		T5/ABC	<	T1/ABCabc	**0.01**	9,9	H				
		T5/ABC	<	T5/ACac	**0.01**	11,2	O				
		T5/ABC	<	T5/ABCabc	**0.01**	18,8	O				
		T5/Aa	<	T5/ABCabc	**0.01**	15,6	O				
		T1/ABC	<	T5/ABab	**0.05**	9,7	H				
		T5/AB	<	T5/ABab	**0.05**	9,7	H				
		T5/ABC	<	T5/ABab	**0.05**	5,6	H				
		T5/Aa	<	T1/ABCabc	**0.05**	6,7	H				
		T5/Aa	<	T5/ACac	**0.05**	8,0	H				
		T5/ABab	<	T5/ABCabc	**0.05**	13,2	O				
		T5/A	<	T5/Aa	0.1	9,0	H				
		T5/AC	<	T5/Aa	0.1	9,0	H				

4 KAIKKI TUTKIMUSYMPÄRISTÖT

Jouk ko	JA/ Virke	SAANTI i		j	α	$x_i - x_j$	SJ	TARKKUUS i		j	α	$x_i - x_j$	SJ
Perus- joukko N = 25	JA	T3/ABC	<	T1/ABC	**0.025**	4,3		(ei merkit-					
		T3/ABC	<	T2/ABC	**0.025**	4,1		seviä eroja)					
		T3/ABC	<	T4/ABC	**0.025**	4,3							
		T3/ABC	<	T5/ABC	**0.025**	7,4	H						
	Virke	T3/ABC	<	T1/ABC	**0.01**	3,2		(ei merkit-					
		T3/ABC	<	T2/ABC	**0.01**	3,1		seviä eroja)					
		T3/ABC	<	T4/ABC	**0.01**	3,3							
		T3/ABC	<	T5/ABC	**0.01**	4,1							
Johdos -osaj. N = 8	JA	T3/ABC	<	T1/ABC	0.1	9,0	H	(ei merkit-					
		T3/ABC	<	T2/ABC	0.1	9,0	H	seviä eroja)					
		T3/ABC	<	T4/ABC	0.1	9,7	H						
		T3/ABC	<	T5/ABC	0.1	15,1	O						
	Virke	T3/ABC	<	T1/ABC	**0.025**	6,3	H	(ei merkit-					
		T3/ABC	<	T2/ABC	**0.025**	6,3	H	seviä eroja)					
		T3/ABC	<	T4/ABC	**0.025**	7,0	H						
		T3/ABC	<	T5/ABC	**0.025**	9,1	H						
Yhdys- sana- osaj. N = 9	JA	T3/ABCabc	<	T1/ABCabc	**0.01**	4,6		T3/ABCabc	>	T1/ABCabc	**0.01**	8,3	H
		T3/ABCabc	<	T2/ABCabc	**0.01**	4,6		T3/ABCabc	>	T2/ABCabc	**0.01**	7,4	H
		T3/ABCabc	<	T4/ABCabc	**0.01**	15,1	H	T4/ABCabc	>	T1/ABCabc	**0.05**	2,5	
		T3/ABCabc	<	T5/ABCabc	**0.01**	14,9	O	T5/ABCabc	>	T1/ABCabc	**0.05**	4,5	
		T1/ABCabc	<	T5/ABCabc	0.1	10,3	O	T3/ABCabc	>	T4/ABCabc	**0.05**	5,8	H
		T2/ABCabc	<	T5/ABCabc	0.1	10,3	O	T3/ABCabc	>	T5/ABCabc	0.1	3,8	
	Virke	T5/ABCabc	>	T1/ABCabc	**0.025**	8,9	H	(ei merkit-					
		T5/ABCabc	>	T2/ABCabc	**0.025**	8,9	H	seviä eroja)					
		T5/ABCabc	>	T3/ABCabc	**0.025**	9,5	H						
		T5/ABCabc	>	T4/ABCabc	**0.05**	8,3	H						
		T3/ABCabc	<	T4/ABCabc	0.1	1,2							

ONGELMAKYSELYJEN OSUMAT

Haun numero	**31**	**32**	**33**	**34**	**35**	**36**	**37a**	**37b**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45a**	**45b**
Perinteinen																	
* osumia yhteensä	**2**	**9**	**50**	**8**	**96**	**33**	**3**	**6**	**14**	**27**	**36**	**9**	**17**	**130**	**8**	**5**	**5**
*** oikein**	**2**	**5**	**49**	**1**	**92**	**31**	**3**	**5**	**7**	**17**	**8**	**4**	**17**	**0**	**8**	**2**	**2**
* hakusanan johdos								**1**	**3**							**3**	**3**
* hakusanalla alkava yhdyssana		**4**								**5**	**14**			**35**			
* hakusanan homografi											**14**	**1**		**95**			
* muu sana, joka alkaa samoin kuin hakusana										**5**		**4**					
* epätarkka operaattori (hakemisto-osoite)			**1**	**7**	**1**	**2**			**4**								
* muu poikkeama (esim. lyhennetty muoto)					**3**												

	31	32	33	34	35	36	37a	37b	38	39	40	41	42	43	44	45a	45b
Finstems, hakusanat suoraan kyselystä																	
* osumia yhteensä	**2**	**9**	**26**	**10**	**54**	**14**	**3**	**5**	**12**	**27**	**36**	**9**	**17**	**135**	**8**	**2**	**6**
*** oikein**	**2**	**5**	**26**	**1**	**54**	**14**	**3**	**5**	**7**	**17**	**8**	**4**	**17**	**0**	**8**	**2**	**2**
* hakusanan johdos									**1**								**4**
* hakusanalla alkava yhdyssana		**4**								**5**	**14**			**40**			
* hakusanan homografi											**14**	**1**		**95**			
* muu sana, joka alkaa samoin kuin hakusana										**5**		**4**					
* epätarkka operaattori (hakemisto-osoite)				**9**					**4**								
* muu poikkeama																	
Finstems, hakusanat korjattu perusmuotoon																	
* osumia yhteensä		**9**	**92**	**28**	**93**	**23**	**3**	**5**	**14**								
*** oikein**		**5**	**49**	**1**	**92**	**21**	**3**	**5**	**7**								
* hakusanan johdos									**3**								
* hakusanalla alkava yhdyssana		**4**															
* hakusanan homografi																	
* muu sana, joka alkaa samoin kuin hakusana																	
* epätarkka operaattori (hakemisto-osoite)			**43**	**27**	**1**	**2**			**4**								
* muu poikkeama																	

	31	32	33	34	35	36	37a	37b	38	39	40	41	42	43	44	45a	45b
Hahmotin, hakusanat suoraan kyselystä																	
* osumia yhteensä	**2**	**9**	**64**	**10**	**93**	**16**	**3**	**5**	**12**	**27**	**36**	**9**	**17**	**137**	**8**	**2**	**6**
*** oikein**	**2**	**5**	**49**	**1**	**92**	**16**	**3**	**5**	**7**	**17**	**8**	**4**	**17**	**0**	**8**	**2**	**2**
* hakusanan johdos									**1**								**4**
* hakusanalla alkava yhdyssana		**4**								**5**	**14**			**42**			
* hakusanan homografi											**14**	**1**		**95**			
* muu sana, joka alkaa samoin kuin hakusana										**5**		**4**					
* epätarkka operaattori (hakemisto-osoite)			**15**	**9**	**1**				**4**								
* muu poikkeama																	
Hahmotin, hakusanat korjattu perusmuotoon																	
* osumia yhteensä		**9**	**92**	**28**	**93**	**34**	**3**	**5**	**14**								
*** oikein**		**5**	**49**	**1**	**92**	**31**	**3**	**5**	**7**								
* hakusanan johdos									**3**								
* hakusanalla alkava yhdyssana		**4**															
* hakusanan homografi																	
* muu sana, joka alkaa samoin kuin hakusana																	
* epätarkka operaattori (hakemisto-osoite)			**43**	**27**	**1**	**3**			**4**								
* muu poikkeama																	

	31	32	33	34	35	36	37a	37b	38	39	40	41	42	43	44	45a	45b
Perusmuotoon palautus																	
* osumia yhteensä	**2**	**5**	**58**	**27**	**93**	**32**	**3**	**5**	**12**	**-**	**166**	**-**	**-**	**95**	[1]**2**	**-**	**-**
*** oikein**	**2**	**5**	**49**	**1**	**92**	**31**	**3**	**5**	**7**		**8**			**0**	**2**		
* hakusanan johdos									**1**								
* hakusanalla alkava yhdyssana											**14**						
* hakusanan homografi											**14**			**95**			
* muu sana, joka alkaa samoin kuin hakusana																	
* epätarkka operaattori (hakemisto-osoite)			**9**	**25**	**1**	**1**			**4**								
* muu poikkeama				**1**							**130**						
Perusmuotoon palautus + seulonta vartaloilla																	
* osumia yhteensä	**2**	**5**	**26**	**8**	**54**	**16**	**3**	**5**	**12**	**27**	**36**	**7**	**17**	**95**	**2**	**2**	**6**
*** oikein**	**2**	**5**	**26**	**1**	**54**	**15**	**3**	**5**	**7**	**17**	**8**	**4**	**17**	**0**	**2**	**2**	**2**
* hakusanan johdos									**1**								**4**
* hakusanalla alkava yhdyssana										**5**	**14**						
* hakusanan homografi											**14**			**95**			
* muu sana, joka alkaa samoin kuin hakusana										**5**		**3**					
* epätarkka operaattori (hakemisto-osoite)				**7**		**1**			**4**								
* muu poikkeama																	

[1] koska lanka-sana oli sanakirjassa, virhettä ei havaittu automaattisesti ja siksi taivutusmuotohakemistossa olevat muodot jäivät löytymättä (jos olisi haettu taivutusvartaloilla tunnistamattomien sananmuotojen hakemistosta, olisivat kaikki 8 relevanttia dokumenttia löytyneet)

	31	32	33	34	35	36	37a	37b	38	39	40	41	42	43	44	45a	45b
Haku kaksoishakemistosta																	
* osumia yhteensä	**2**	**5**	**26**	**8**	**54**	**16**	**3**	**5**	**12**	**27**	**36**	**9**	**17**	**95**	[2)]**2**	**2**	**6**
*** oikein**	**2**	**5**	**26**	**1**	**54**	**15**	**3**	**5**	**7**	**17**	**8**	**4**	**17**	**0**	**2**	**2**	**2**
* hakusanan johdos									**1**								**4**
* hakusanalla alkava yhdyssana										**5**	**14**						
* hakusanan homografi											**14**	**1**		**95**			
* muu sana, joka alkaa samoin kuin hakusana										**5**		**4**					
* epätarkka operaattori (hakemisto-osoite)				**7**		**1**			**4**								
* muu poikkeama																	

[2)] koska lanka-sana oli sanakirjassa, virhettä ei havaittu automaattisesti ja
siksi taivutusmuotohakemistossa olevat muodot jäivät löytymättä
(jos olisi haettu taivutusvartaloilla perinteisestä hakemistosta, olisivat kaikki 8 relevanttia dokumenttia löytyneet)