Jaana Kristensen

Hakutesauruksen vaikutus sanahakuihin

Kristensen, Jaana, Hakutesauruksen vaikutus sanahakuihin [The effectiveness of a search-aid thesaurus in query expansion of free-text searches]. Kirjastotiede ja informatiikka 12 (3): 95–104, 1993.

Typical problem in free-text searching of full-text databases is the selection of accurate search words. This paper discusses the effectiveness of a thesaurus as a source for search words in query expansion of free-text searches. A test was conducted in a large full-text database containing newspaper articles. End-users' queries were searched in five distinct modes: basic search, synonym search, narrower term search, related term search and union of all previous searches. The basic searches contained only words included in the original query statements. In the other modes, the basic searches were expanded by disjunction of the terms given by the thesaurus. The last search mode included the basic words and all the terms used in the previous searches. The searches were analyzed in terms of relative recall and precision; relative recall was based on the recall of the union search which was given a value of 100 %. On the average, relative recall was 47,2 % in the basic search compared with 100% in the union search; the average value of precision decreased only from 62,5% in the basic search to 51,2% in the union search.

Address: University of Tampere, Department of Information Studies, P.O.Box 607, SF-33101 Tampere, Finland.

1. Hakusanojen valinnan ongelma sanahaussa

Tekstitietokannat sisältävät sanallista informaatiota, usein dokumenttien koko tekstin. Tekstimuotoisen informaation tallennuksen ja haun ongelmat liittyvät osittain kielen ominaisuuksiin. Luonnolliselle kielelle tyypillisiä piirteitä ovat ilmaisujen ja niiden sävyjen rikkaus sekä ilmaisujen merkitysten kontekstisidonnaisuus. Samaan asiaan tai ilmiöön voidaan viitata monilla eri ilmauksilla, jolloin eri muotoisten ilmausten merkityssisällöt ovat samat tai osittain päällekkäiset (yksinkertaistaen synonymia). Vastakkainen kielen piirre on monimerkityksisyys (yksinkertaistaen homonymia), jossa samanmuotoisten ilmausten merkitykset eroavat toisistaan. Kun tiedonhaku kohdistuu dokumenttien teksteihin, aiheuttavat synonymia ja homonymia hakujen epätarkkuutta ja saantivirheitä, koska eri kirjoittajat käyttävät kieltä eri tavoin, eivätkä tiedonhakijoiden ja kirjoittajienkaan sanavalinnat osu yksiin (Furnas ym. 1987; Gomez ym. 1990). Kontrolloitua sanastoa käyttävä dokumenttien sisällönkuvailu pyrkii yhdenmukaistamaan tiedon tallennusta ja hakua, mutta sekään ei ole ongelmatonta, sillä indeksoinnin tulos on tekstin yksi tulkinta, joka sulkee muut mahdolliset tulkinnat pois. Lisäksi indeksointi on kallista.

Tiedonhakijan työtä voidaan helpottaa tarjoamalla erilaisia sanastoja ja muita apuneuvoja hakusanojen valintaan. Tällaisia ovat esimerkiksi synonyymisanastot, tesaurukset, käänteistiedoston selausmahdollisuus ja aikaisempien hakuprofiilien

tallentaminen. Monet näistä ovatkin käytössä yleisissä tiedonhakujärjestelmissä, mutta kaikki kuvailutermejä sisältävät tietokannatkaan eivät tarjoa suorasaantista tesaurusta hakijan avuksi. (Aitchison & Gilchrist 1987; Lancaster 1986; Piternick 1984; Tenopir & Ro 1990.) Tesauruksia käytetäänkin ennen kaikkea tiedon tallennusvaiheessa ja vasta toissijaisesti haussa. Tesaurus voidaan kuitenkin ottaa hakukäyttöön ympäristöissä, joissa dokumenttien sisältöä ei ole kuvailtu kontrolloidulla sanastolla (Lancaster 1972). Tällöin sanahakua tuetaan esittämällä hakijalle hänen hakuaiheeseensa liittyviä sanoja, jotka on järjestetty keskinäisten semanttisten suhteidensa perusteella. Tällaisia suhteita esittävä *hakutesaurus* eroaa tavanomaisesta tesauruksesta siten, että se antaa hakijalle monia vaihtoehtoja hakusanoiksi eikä ohjaa useasta synonyymisesta ilmauksesta yhteen kuvailutermiin. Hakutesauruksessa ekvivalenttien ja assosiatiivisten suhteiden runsaus on keskeistä. Hakutesauruksen tarkoitus on osoittaa hakijalle valikoima tietyn aihealueen käsitteiden mahdollisista nimistä sekä käsitteiden välisistä suhteista. Ajatus hakujen laajentamisesta tesauruksen avulla ei ole uusi, mutta sen testaus ja soveltaminen on jäänyt vähäiseksi. Lisensiaatintutkimukseni käsitteli hakutesauruksen käytön vaikutusta sanahauissa (Kristensen 1992). Tässä artikkelissa esittelen tutkimuksen keskeiset tulokset.

Tesauruksen käyttöä hakujen laajentamisessa on testattu aikaisemmin vektorimalliin perustuvissa hakujärjestelmissä ja käytetyt tietokannat ovat olleet pieniä testikantoja (esim. Fox 1980; Fox 1988; Lu 1990; Wang ym. 1985). Boolen logiikkaa käyttävässä ympäristössä hakutesaurusta on testattu tämän työn esitutkimuksessa, jonka tulokset osoittivat hakutesauruksen käytön kasvattavan sanahakujen suhteellista saantia ilman tarkkuuden suurta heikkenemistä (Kristensen & Järvelin 1990). Työtä jatkettiin niin, että testattiin laajempaa hakutesaurusta suuremmassa hakuympäristössä kuin esitutkimuksessa. Myös testattavien hakulaajennostyyppien määrää lisättiin. Tässä artikkelissa selitetään uuden tutkimuksen sanavalintoja sekä esitellään sen tutkimusasetelma, tulokset ja johtopäätökset.

2. Sanoista, termeistä ja käsitteistä

Ilmaus on yksi sana tai useamman sanan muodostama kokonaisuus, kuten sanaliitto. Sana on luonnollisen kielen yksikkö, jonka merkitys elää kielessä (ks. Karlsson 1980, 105–110). Termi on luonnollisesta kielestä poimittu sana tai sanaliitto, jonka merkitys on rajattu ja määritelty (oppi- tai ammattisana, ks. Haarala 1981, 15). Termi voi olla myös muu merkki kuin sana, mutta tässä artikkelissa termit ovat sanoja. Hakusana ja hakutermi – sanoja käytetään usein samaa tarkoittavina, eli niillä viitataan ilmauksiin tai termeihin, joilla hakuprofiilissa kuvataan hakuaihetta. Tässä tekstissä tarkoitan hakusanalla joko luonnollisen kielen sanaa tai kontrolloidusta sanastosta poimittua termiä, hakutermillä tarkoitan vain jälkimmäistä. Kuvailutermillä tarkoitan dokumenttien sisällönkuvailussa käytettyjä termejä.

Sanahaulla tarkoitan tiedonhakua, jossa hakusanat voidaan poimia mistä tahansa lähteestä. Painotus on silloin hakusanojen valinnassa. Jos halutaan korostaa haun kohteena olevan nimenomaan dokumentin koko tallennettu teksti, voidaan puhua tekstihausta (ks. Sormunen & Alkula 1990, 8). Aikaisemmin olen käyttänyt muotoja vapaatekstihaku ja vapaasanahaku, jotka ovat käännösväännöksiä englannista. Jos sanan ja termin välillä tehdään edellä mainittu merkitysero, voidaan etuliite vapaa jättää tarpeettomana pois. Sanahakua varten rakennettu hakutesaurus sisältää sanoja ja sanaliittoja. Se ennemminkin kuvailee kuin kontrolloi kieltä. Toisaalta hakutesaurus käyttää tesaurusterminologiaa (kuten rinnakkaistermit, suppeammat termit), minkä vuoksi sen esittämät suhteet vaikuttavat normatiivisilta, vaikka eivät aina teksteissä toteutuisikaan. Yksinkertaisuuden vuoksi puhun hakutesauruksen ilmauksista termeinä.

Kielen ja maailman suhteesta ei vallitse yksimielisyyttä. Suhteen kuvauksissa esiintyy kuitenkin usein *käsite*. Käsitteet on määritelty muun muassa ajattelun abstraktioiksi (mental objects), jotka perustuvat maailmasta aistien välityksellä saatuihin havaintoihin (esim. Allan 1986; Baldinger 1980; Karlsson 1980). Määrittelyyn sisältyy epäilemättä ongelmia, mutta käsite on tarpeellinen apuväline. Kieli nimeää käsitteet. Ajateltiinpa käsitteitä kieleen kuuluvina tai kielen ulkoisina, niihin viitataan sanoilla tai termeillä. Yhtä käsitettä voidaan kuvata useammalla kuin yhdellä ilmauksella, ja toisaalta yhdellä ilmauksella voidaan viitata useampaan kuin yhteen käsitteeseen. Tämän takia käsitteiden ja niiden nimien (ilmauksia tai termejä) tasot on pidettävä erillään (Soergel 1985, 271).

3. Tutkimuksen aineisto ja menetelmät

Tutkimuksen tarkoituksena oli selvittää, miten sanahakujen laajentaminen hakutesauruksen avulla vaikuttaa hakujen saantiin ja tarkkuuteen. Työssä tarkasteltiin hakutesauruksen kokonaisvaikutusta ja sen erityyppisten termiryhmien vaikutusta hakutuloksiin. Lisäksi testattiin hakutesauruksen vaikutusta tuloksiin käytettäessä läheisyysoperaattoreita Boolen JA-operaattorin asemasta.

3.1 Tutkimusympäristö

Koeympäristöksi valittiin suuri tekstitietokanta, jossa dokumentteja ei ole kuvailtu kontrolloitua sanastoa käyttäen. Sanahaun ongelmat ovat tällaisessa ympäristössä ilmeisempiä ja tutkimustulokset luotettavampia kuin pienissä testitietokannoissa (Blair & Maron 1985). Testihaut tehtiin *Aamulehden* tekstitietokannasta, joka sisälsi sanomalehtiartikkeleita. Tietokannassa oli noin 225 000 artikkelia ja artikkelien keskimääräinen pituus oli 5 000 merkkiä (n. 2 A4-sivua). Tietokannan hallinta- ja hakuohjelmistona *Aamulehden* tekstitietokannassa on BASIS (ks. McDonald 1984), jossa haku perustuu Boolen logiikan ja käänteistiedostorakenteen käyttöön. Tietokannan käänteistiedosto on rakennettu sanojen perusmuodoista, joka nomineilla on yksikön nominatiivi, verbeillä I infinitiivi. Nämä perusmuodot on saatu artikkeleiden teksteistä suomenkielisten sanojen taivutusmuotoja perusmuotoihin palauttava MORFO-ohjelman avulla (ks. Alkula & Honkela 1992, 26). Tämän vuoksi sanojen taipumista ei tarvitse ottaa huomioon hakujen muotoilussa. Lisäksi ohjelma jakaa yhdyssanat osiin niin, että esimerkiksi hakusanalla *investointi* saadaan tekstistä osumiksi myös yhdyssanat *teollisuusinvestointi, pääomainvestointi, investointivaltuuskunta,* jne. Menettely parantaa saantia, mutta heikentää tarkkuutta.

3.2 Hakutesaurus

Hakutesauruksen ilmaisemat semanttiset suhteet ovat samanlaisia kuin tavanomaisessa tesauruksessa, mutta käyttötapa on päinvastainen. Hakutesauruksen sanaston täytyy vastata haettavien dokumenttien kieltä, muuten siitä ei ole hyötyä hakujen laajentamisessa. Lisäksi hakutesauruksen täytyy esittää enemmän ekvivalentteja termejä kuin tavallisen tesauruksen, koska synonyymiset ilmaukset ovat oleellisia haun laajentamisessa. Koska tällaista sanastoa ei ollut valmiina, tutkimusta varten rakennettiin sanasto, jonka aiheina olivat ympäristö ja talous. Hakutesaurus ei kattanut aihealueitaan perusteellisesti, vaan termit kuvasivat sanomalehdissä usein toistuvia uutisaiheita. Kaikkiaan tesaurukseen kertyi 1573 termiä, jotka kuvaavat 1011 käsitettä, kun eri käsitteiksi lasketaan kaikki termit, jotka eivät ole keskenään synonyymejä.

Hakutesaurus laadittiin ISO:n standardin 2788/1986 ohjeita seuraten. Hakutesaurus eroaa kuitenkin indeksointikielestä siten, että edellinen pyrkii vain kuvailemaan kieltä, kun taas jälkimmäinen on normatiivista pyrkiessään tallennuksen yhdenmukaisuuteen. Tämän eron vuoksi ISO:n ohjeista poikettiin seuraavasti:

- Hakutesauruksessa kaikki termit ovat mahdollisia hakutermejä, siksi kuvailu- ja ohjaustermejä ei ole eroteltu.
- Termit eivät ole sanaluokaltaan pelkästään substantiiveja vaan myös adjektiiveja ja verbejä. Nämä sananluokat on otettu mukaan, koska on kysymys tekstitietokannasta, jossa eri sanaluokkien sanoja voidaan hakea, ja koska eri sanaluokkien ominaisuuksista hakusanoina ei vielä tiedetä paljoakaan.
- Termit ovat yksikössä, koska tietokannan käänteistiedostoon on tallennettu sanojen perusmuodot ja hakija käyttää yksikkömuotoja.

Tavanomaisissa tesauruksissa ekvivalenssisuhde tulkitaan hyvin tiukasti merkitykseltään kahden täysin yhtenevän termin väliseksi suhteeksi. Absoluuttinen synonymia on kuitenkin harvinaista, sillä usein sanoilla on ainakin sävyero. Siksi synonymia ymmärretään hakutesauruksessa väljemmin: Samasta kannasta lähtöisin olevat verbi, adjektiivi ja substantiivi voivat olla synonyymeja (*ympäristöpolitiikka* ja *ympäristöpoliittinen* tai *investointi* ja *investoida*) tai synonyymien merkitys voi olla vain osittain yhteinen (*erottaa* ja *irtisanoa*). Menettelyn tarkoituksena oli sopeuttaa ekvivalenssisuhteet vastaamaan sanahakua sekä lisätä synonyymien määrää.

3.3 Hakukysymykset ja hakuprofiilit

Tietokannan käyttäjiltä kerättiin kirjallisia hakukysymyksiä. Testin 30 hakukysymyksestä 21 ke-

rättiin *Aamulehden* toimittajilta ja 9 kysymystä eräältä lehdistöseurantaa tekevältä tietopalveluyritykseltä. Kysymykset olivat toimittajien ja tietopalveluyrityksen asiakkaiden aihehakuja talouden ja ympäristön alalta. Kysymykset kerättiin kirjallisina, jotta aiheet ilmaistaisiin tiedontarvitsijoiden omilla sanoilla. Tämä oli tarpeen, koska tarkoituksena oli mallintaa tilannetta, jossa tiedontarvitsijan antamista hakusanoista lähtien hakua laajennetaan automaattisesti. Kysymyksistä valittiin ne, joiden sanoista ainakin osa löytyi hakutesauruksesta ja joiden avulla tehdyn sanahaun tulosjoukossa oli vähintään yksi relevantti artikkeli.

Jokaisesta hakukysymyksestä tehtiin perushaku analysoimalla kysymyksen peruskäsitteet ja ilmaisemalla ne hakukysymyksen sanoin. Esimerkiksi hakukysymyksen "Yhtyneiden Paperitehtaiden, Metsäliiton ja Veitsiluodon metsäteollisuuden investoinnit" peruskäsitteitä ovat Yhtyneet Paperitehtaat, Metsäliitto, Veitsiluoto, metsäteollisuus ja investointi. Haut noudattivat Boolen logiikkaa, leikkaavat käsitteet yhdistettiin JA-operaattorilla, vaihtoehtoiset käsitteet tai käsitteiden vaihtoehtoiset nimet TAI-operaattorilla. Peruskäsitteet olivat yleensä yksittäisiä sanoja, erisnimissä esiintyi sanaliittoja.

Tämän jälkeen hakuja laajennettiin hakutesauruksen eri termiryhmillä, eli hakutesauruksesta liitettiin hakuprofiiliin termejä, joita se antoi perushaun hakusanoille. Laajennoksia oli neljän tyyppisiä: (1) synonyymien lisääminen perushakuun eli synonyymihaku; (2) suppeampien termien lisääminen perushakuun eli hierarkkinen haku; (3) rinnakkaistermien lisääminen perushakuun eli rinnakkaistermihaku; (4) kaikkien edellisten lisääminen perushakuun eli laajin haku (Ks. Kuvio 1). Viimeisen eli laajimman haun tulosjoukko sisälsi muiden hakujen tulosjoukot. Hakutyypit merkitään **P**, **S**, **H**, **R**, **L** ja niitä vastaavat tulosjoukot ***P***, ***S***, ***H***, ***R***, ***L***. Silloin

$$\boldsymbol{S} \supseteq \boldsymbol{P}\,;\, \boldsymbol{H} \supseteq \boldsymbol{P}\,;\, \boldsymbol{R} \supseteq \boldsymbol{P} \quad \text{ja}$$
$$\boldsymbol{L} \supseteq (\boldsymbol{S} \cup \boldsymbol{H} \cup \boldsymbol{R}).$$

Kysymyksen esittänyt toimittaja arvioi laajimman haun tulosjoukon artikkeleiden relevanssin relevantti/epärelevantti -arviona. Tietopalvelukysymysten osalta relevanssiarviot teki yksi artikkeleita kysytyistä aiheista etsivä työntekijä. Mikäli tulosjoukko oli liian suuri (>100 artikkelia), rajattiin haku artikkeleihin vuosilta 1990–91 (9 hakukysymystä). Mikäli aikarajauskaan ei pienentänyt tulosjoukkoa tarpeeksi, tehtiin satunnaisotanta (6 hakukysymystä). Otannan jälkeen 30 hakukysymyksen laajimpien hakujen tulosjoukot sisälsivät 1384 artikkelia, joista 661 oli relevanttia ja 723 epärelevanttia.

4. Tulokset

Kaikista hakukysymyksistä ei voitu tehdä kaikkia laajennostyyppejä, koska hakutesauruksesta ei löytynyt tarvittavia termejä (esimerkiksi muutaman hakukysymyksen perushaun sanoille ei ollut lainkaan synonyymeja). Perushaku ja laajin haku, jossa olivat kaikki aikaisempien mahdollisten hakujen hakusanat, tehtiin kaikista 30 hakukysymyksestä. Kaikki laajennokset tehtiin 14 hakukysymyksestä. Tuloksissa esitetään siis 30 hakukysymyksen osalta tesauruksen kokonaisvaikutus, mutta vain 14 hakukysymyksen osa-aineistosta kaikkien eri hakutyyppien tehokkuus.

4.1. Hakujen suhteellinen saanti ja tarkkuus

Suurissa tietokannoissa absoluuttista saantia on mahdotonta selvittää, siksi tutkimuksissa käytetäänkin usein saannin arviota tai suhteellista saantia (Lancaster 1968). Tässä tutkimuksessa todellista saantia ei edes tarvittu, koska eri hakutyyppejä haluttiin vertailla keskenään. Relevanssiarvioiden perusteella laskettiin jokaisen hakukysymyksen kaikille hauille suhteellinen saanti ja tarkkuus. Laajimman haun suhteellinen saanti oli 100 %, koska se oli saannin laskemisen perusta. Muiden hakutyyppien saanti laskettiin laajimman haun relevanttien artikkeleiden perusteella. Keskiarvot saatiin laskemalla ensin jokaiselle haulle erikseen suhteellinen saanti ja tarkkuus, sen jälkeen laskettiin kaikkien hakujen keskiarvot kummastakin (macro-average) eli saannin ja tarkkuuden keskiarvot määritellään seuraavasti:

$$Ss = [\Sigma_{i=1\ldots n}(r_i / t_i)] * 100 / n$$
$$Ta = [\Sigma_{i=1\ldots n}(r_i / d_i)] * 100 / n$$

jossa r_i on saatujen relevanttien dokumenttien määrä, t_i kaikkien relevanttien dokumenttien määrä, d_i kaikkien saatujen dokumenttien määrä i :nnessä haussa, n on hakukysymysten määrä ja *Ss* ja *Ta* ovat saannin ja tarkkuuden keskiarvot (Tague-Sutcliffe 1992, 483). Tulokset ovat painottamattomia keskiarvoja.

Perushaun suhteellisen saannin keskiarvo oli 47,2 prosenttia. Perushaun ja laajimman haun tarkkuuden

Hakukysymys: Vaatetusteollisuuden konkurssit ja irtisanomiset

P — Perushaku (relevantteja 11; epärelevantteja 8; suhteellinen saanti 45,8%; tarkkuus 57,9%)

VAATETUSTEOLLISUUS
JA
(**KONKURSSI** TAI **IRTISANOMINEN**)

S — Synonyymihaku (relevantteja 13; epärelevantteja 10; suhteellinen saanti 54,2%; tarkkuus 56,5%)

(**VAATETUSTEOLLISUUS** TAI **VAATETEOLLISUUS**)
JA
(**IRTISANOMINEN** TAI **IRTISANOA** TAI **EROTTAA** TAI **EROTTAMINEN** TAI **KONKURSSI**)

H — Hierarkkinen haku (relevantteja 12; epärelevantteja 8; suhteellinen saanti 50%; tarkkuus 60%)

(**VAATETUSTEOLLISUUS** TAI **VALMISVAATETEOLLISUUS** TAI **SUKKATEOLLISUUS** TAI **TRIKOOTEOLLISUUS**)
JA
(**IRTISANOMINEN** TAI **KONKURSSI**)

R — Rinnakkaistermihaku (relevantteja 20; epärelevantteja 29; suhteellinen saanti 83,3%; tarkkuus 40,8%)

(**VAATETUSTEOLLISUUS** TAI **TEKSTIILITEOLLISUUS**)
JA
(**IRTISANOMINEN** TAI **KILOMETRITEHDAS** TAI **LOMAUTTAA** TAI **LOMAUTTAMINEN** TAI **LOMAUTUS** TAI **PAKKOLOMA** TAI **SANEERAAMINEN** TAI **SANEERATA** TAI **SANEERAUS** TAI **SANEERAUSTOIMI** TAI **TERVEYTTÄÄ** TAI **TERVEYTYS** TAI **TYÖTTÖMYYS** TAI **TYÖTÖN** TAI **YT+NEUVOTTELU** TAI **KONKURSSI**)

L — Laajin haku (relevantteja 24; epärelevantteja 33; suhteellinen saanti 100%; tarkkuus 42,1%)

(**VAATETUSTEOLLISUUS** TAI **VAATETEOLLISUUS** TAI **VALMISVAATETEOLLISUUS** TAI **SUKKATEOLLISUUS** TAI **TRIKOOTEOLLISUUS** TAI **TEKSTIILITEOLLISUUS**)
JA
(**IRTISANOMINEN** TAI **IRTISANOA** TAI **EROTTAA** TAI **EROTTAMINEN** TAI **KILOMETRITEHDAS** TAI **LOMAUTTAA** TAI **LOMAUTTAMINEN** TAI **LOMAUTUS** TAI **PAKKOLOMA** TAI **SANEERAAMINEN** TAI **SANEERATA** TAI **SANEERAUS** TAI **SANEERAUSTOIMI** TAI **TERVEYTTÄÄ** TAI **TERVEYTYS** TAI **TYÖTTÖMYYS** TAI **TYÖTÖN** TAI **YT+NEUVOTTELU** TAI **KONKURSSI**)

Merkkien selitykset:
IRTISANOA = hakusana/lisätty hakutermi
JA = Boolen operaattori
TAI = Boolen operaattori
+ = hakutermi haettu sanaliittona läheisyysoperaattoria käyttäen

Kuvio 1. Esimerkki yhden hakukysymyksen eri hakutyypeistä

erot olivat pienemmät kuin niiden suhteellisen saannin erot. Vaikka saannin kasvu johti tarkkuuden heikkenemiseen, oli laajimman haun tarkkuus keskimäärin vain kymmenisen prosenttiyksikköä pienempi kuin perushaun (Taulukko 1).

Perushaun ja laajimman haun erojen tilastollisen merkitsevyyden testaamiseen käytettiin epäparametristä Wilcoxonin testiä (Siegel & Castellan, 1988). Suhteellisen saannin osalta erot olivat tilastollisesti erittäin merkitseviä ($p>0,0001$). Tarkkuuden erot olivat merkitseviä ($p>0,01$), joskaan nollahypoteesia (tarkkuuden arvo ei muutu) ei voitu hylätä yhtä pienellä riskillä.

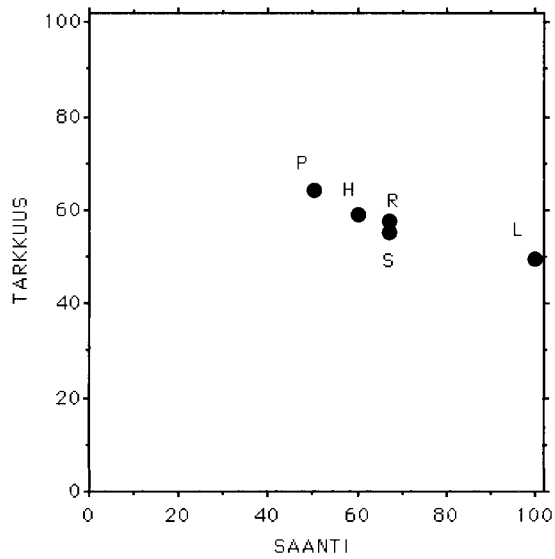
Taulukko 1. Perushaun ja laajimman haun suhteellisen saannin ja tarkkuuden keskiarvot (N=30).

Hakutyyppi	Suhteellinen saanti %	Tarkkuus %
Perushaku	47,2	62,5*
Laajin haku	100,0	51,2

* N = 27, kolmessa hakukysymyksessä perushaku ei antanut tuloksia, tarkkuuden arvoa ei voitu laskea

Neljäntoista hakukysymyksen osa-aineistossa, josta tehtiin kaikki hakutyypit, perushaun suhteellinen saanti oli 50,3% ja tarkkuus 64,4%; laajimman haun vastaavasti 100%/49,5%. Synonyymihaun, hierarkkisen haun ja rinnakkaistermihaun tulokset eivät poikenneet toisistaan kovin paljon (Kuvio 2). Hierarkkisen haun suhteellinen saanti oli huonoin (60%) ja tarkkuus paras (59,1%), synonyymihaku oli saanniltaan (67,2%) ja tarkkuudeltaan (55,2%) vain vähän rinnakkaistermihakua heikompi (67,3%/57,4%). Laajimman haun ero muihin hakuihin viittaa siihen, että eri termityypeillä saadut tulosjoukot eivät olleet identtiset.

Kun kaikkien viiden eri hakutyypin saannin ja tarkkuuden eroja vertailtiin keskenään, käytettiin tilastollisen merkitsevyyden selvittämisessä Friedmanin testiä (Kinnucan ym. 1987; Siegel & Castellan 1988). Tulos oli samanlainen kuin edelläkin: suhteellisen saannin erot olivat erittäin merkitseviä ($p<0,0001$) ja tarkkuuden erot merkitseviä ($p<0,01$). (Testeistä lähemmin ks. Kristensen 1992.)



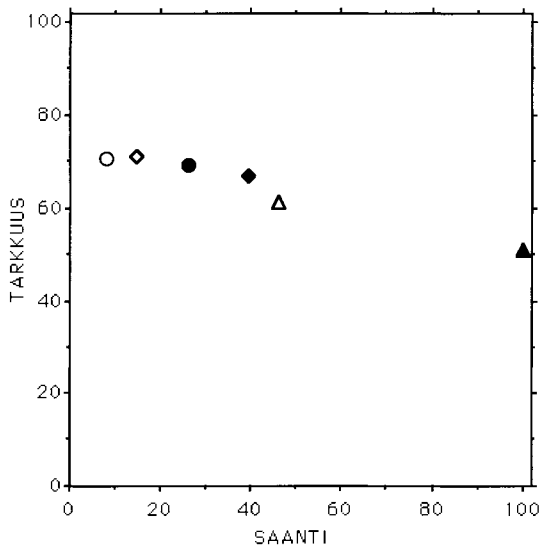
Kuvio 2. Viiden eri hakutyypin suhteellisen saannin ja tarkkuuden pisteparvi neljäntoista hakukysymyksen osa-aineistosta. Symbolit: P = perushaku; H = hierarkkinen haku; R = rinnakkaistermihaku; S = synonyymihaku; L = laajin haku.

4.2 Hakutesauruksen vaikutus läheisyysoperaattoreita käytettäessä

Tekstitiedonhaussa JA-operaattorin käyttö aiheuttaa usein epätarkkuutta. JA-operaattori edellyttää termien esiintyvän samassa dokumentissa. Jos teksti on pitkä, voivat ehdon täyttävät termit olla hyvinkin etäällä toisistaan ja kuulua eri yhteyksiin. Läheisyysoperaattoreiden käytöllä pystytään parantamaan tarkkuutta. Niiden avulla voidaan (1) rajata tekstin rakenneosa (kappale, lause), jossa hakusanan on esiinnyttävä; (2) määrätä kahden tai useamman hakusanan keskinäinen etäisyys; (3) määrätä kahden tai useamman hakusanan esiintymisjärjestys. (Keen 1991; Keen 1992; Tenopir & Shu 1989.) Tässä tutkimuksessa saaduista tulosjoukoista tarkastettiin jälkikäteen, mitä tulosjoukkoihin olisi kuulunut, jos koko dokumenttiin kohdistuvat JA-operaattorit (DOKUMENTTI; ks. Kuvio 3) olisi korvattu (a) leikkaavat käsitteet yhteen virkkeeseen rajoittavalla läheisyysoperaattorilla (VIRKE); (b) leikkaavat käsitteet yhteen kappaleeseen rajoittavalla läheisyysoperaattorilla (KAPPALE). Suhteellisen saannin ja tarkkuuden arvot las-

kettiin perushaun ja laajimman haun eri operaattoreilla saaduille tulosjoukoille. Vertailu perustuu 26 hakukysymyksen aineistoon, koska neljän hakukysymyksen hakuprofiilissa ei käytetty JA-operaattoria.

Läheisyysoperaattoreiden käyttö paransi tarkkuutta mutta heikensi saantia huomattavasti molemmissa hakutyypeissä (Kuvio 3). Hakukysymyksen leikkaavien käsitteiden rajaaminen yhden virkkeen sisään on hyvin tiukka ehto, joten saannin väheneminen oli odotettua. Myös artikkeleiden kappaleet olivat lyhyitä, mikä osaltaan selittää tulosta. Hakutesauruksen vaikutus tuli kuitenkin esiin verrattaessa hakutyyppejä keskenään: mitä tiukempaa läheisyysoperaattoria käytettiin, sitä enemmän hakutesauruksesta oli hyötyä. Rajattaessa käsitteet



● Laajin haku/VIRKE

○ Perushaku/VIRKE

◆ Laajin haku/KAPPALE

◇ Perushaku/KAPPALE

▲ Laajin haku/DOKUMENTTI

△ Perushaku/DOKUMENTTI

Kuvio 3. Laajimman haun ja perushaun suhteellisen saannin ja tarkkuuden pisteparvi eri operaattoreita käytettäessä.

virkkeeseen oli laajimman haun suhteellinen saanti 3,3 kertaa suurempi kuin perushaussa (26,1% vs 7,9%). **Kappale-** ja **dokumentti**rajauksissa laajimman haun saanti oli 2,7 ja 2,2 kertaa suurempi kuin perushaun (laajin haku: 39,4% ja 100% vs perushaku: 14,5% ja 45,9%). Tarkkuudessa ero hakutyyppien välillä pieneni mitä tiukempaa operaattoria käytettiin. Laajimman haun tarkkuus oli 69,4% ja perushaun 70,7% käytettäessä **virke**rajausta; **kappale**rajausta käytettäessä vastaavat luvut olivat 66,7% ja 71%; **JA**-operaattoria käytettäessä 51,2% ja 61,1%. Läheisyysoperaattoreiden käyttö parantaa laajimman haun tarkkuutta enemmän kuin perushaun. Wilcoxonin testi osoitti, että erot suhteellisessa saannissa perushaun ja laajimman haun välillä olivat erittäin merkitsevät riippumatta käytetystä operaattorista ($p < 0.001$); tarkkuuden ero hakutyyppien välillä oli merkitsevä vain käytettäessä JA-operaattoria ($p < 0.05$).

4.3 Erityyppisten termien tehokkuus hakutermeinä

Eri laajennostyyppien suhteellisen saannin arvoissa oli jonkin verran eroja, mikä merkitsee sitä, että erityyppisten termien vaikutus on erilainen. Jokaiselle lisätylle hakutermille laskettiin, kuinka monta relevanttia ja epärelevanttia uutta artikkelia se keskimäärin toi tulosjoukkoon. Seuraavassa tarkastellaan lisättyjen hakutermien tehokkuutta niiden antamien lisätulosten perusteella. Hakutyypin **T**, **T** ∈ {**S**, **H**, **R**, **L**} lisätulokset T_e määritellään seuraavasti, kun T on hakutyypin **T** tulosjoukko ja P on kyseessä olevan hakukysymyksen perushaun **P** tulosjoukko:

$$T_e = T - P$$

Erityyppisten termien tehokkuutta hakutermeinä voidaan vertailla jakamalla kunkin hakutyypin lisätulokset haussa lisättyjen termien määrällä, eli:

$$\Sigma_{i=1...n} (|T\ e_i| / |Wt_i|) / n$$

jossa $Te_1,..., Te_n$ ovat hakutyypin **T**, **T** ∈ {**S**, **H**, **R**} lisätulokset hakukysymyksille i = 1...n ja $Wt_1,..., Wt_n$ on hakutyypin **T**, **T** ∈ {**S**, **H**, **R**} hakuprofiiliin hakukysymyksissä i = 1...n liitettyjen termien joukko, ja n on niiden hakukysymysten määrä, jossa hakutyypillä oli lisätuloksia. Laajin haku jätettiin pois tarkastelusta, koska sen hakuprofiili ei sisällä uusia hakutermejä.

Synonyymeja lisättiin perushakuun kaikkiaan 20 hakukysymyksessä, keskimäärin 3,7 kappaletta

haussa ja kaikkiaan 74 kappaletta. Jokainen synonyymi toi tulosjoukkoon keskimäärin 2,5 lisäartikkelia, joista 1,3 oli relevantteja. Suppeampia termejä lisättiin 25 hakukysymykseen kaikkiaan 374, keskimäärin 15 kappaletta hakua kohti. Kaikkien saatujen ja saatujen relevanttien artikkeleiden osuudet yhtä suppeampaa termiä kohti olivat 0,4 ja 0,2. Rinnakkaistermejä lisättiin keskimäärin 8 hakua kohti eli 29 hakukysymykseen kaikkiaan 232 kappaletta. Keskimäärin jokainen antoi lisätuloksena 3,6 artikkelia, joista 1,6 oli relevantteja. (Taulukko 2.) Rinnakkaistermit antoivat eniten sekä relevantteja että epärelevantteja lisätuloksia, mutta synonyymien lisätuloksissa oli suhteellisesti eniten relevantteja artikkeleita. Tulos poikkesi hieman hakujen tarkkuuden mukaisesta järjestyksestä, jossa hierarkkinen haku oli tarkin ja synonyymihaku epätarkin. Pieni ero selittyy sillä, että tarkkuuden arvot perustuivat eri osa-aineistoon (14 yhteistä hakukysymystä).

Taulukko 2. Kaikkien ja relevanttien lisätulosartikkeleiden keskimääräinen osuus lisättyjä termejä kohti

Artikkelit / hakutermi	Kaikki saadut lisäartikkelit	Saadut relevantit lisäartikkelit
Synonyymit (N = 20)	2,5	1,3
Suppeammat termit (N = 25)	0,4	0,2
Rinnakkaistermit (N = 29)	3,6	1,6

N = hakukysymysten määrä, joissa ao. termejä lisättiin.

Substantiivit vallitsevat hakutermeinä, mikä epäilemättä johtuu niiden nimeävästä luonteesta. Jackson (1983) mainitsee myös muiden sanaluokkien käyttökelpoisuudesta. Koska hakutesauruksessa oli verbejä ja adjektiiveja, myös niiden tehokkuutta tarkasteltiin selvittämällä miten synonyymien lisätulokset jakaantuivat eri sanaluokkiin kuuluvien synonyymien kesken. Synonyymit valittiin, koska niiden joukossa esiintyi eniten verbejä. Kaikista lisätyistä 74 synonyymista oli 62 substantiiveja, 11 verbejä ja 1 adjektiivi[1]. Taulukosta 3 käy ilmi, että lisätulosten määrä verbiä kohti oli suurempi kuin substantiivia kohti ja näistä lisätuloksista suurempi osa oli relevantteja. Muiden kuin substantiivien määrä hakutermien joukossa on niin vähäinen, että tulokset ovat lähinnä viitteellisiä.

Taulukko 3. Kaikkien ja relevanttien artikkelien keskimääräinen osuus eri sanaluokkien synonyymeja kohti.

Artikkelit / hakutermi	Kaikki saadut artikkelit	Saadut relevantit artikkelit
Substantiivit (N =19)	3,0	1,4
Verbit (N = 7)	5,1	4,6
Adjektiivit (N = 1)	1,0	1,0

N = hakukysymysten määrä, joissa ao. termejä lisättiin.

5. Johtopäätökset

Tutkimuksessa testattiin miten suuresta tekstitietokannasta tehtyjen sanahakujen saanti ja tarkkuus muuttuivat, kun hakuja laajennettiin hakutesauruksesta poimituilla lisätermeillä. Osoittautui, että hakutesauruksen kokonaisvaikutus sanahakujen saannin paranemiseen oli huomattava: Saanti kasvoi lähes kaksinkertaiseksi perushausta (47,2%) laajimpaan hakuun (100%). Tarkkuus muuttui vähemmän, perushaun 62,5 prosentista laajimman haun 51,2 prosenttiin. Perushaussa käytettiin hakukysymyksen sanoja, laajimmassa haussa käytettiin perushaun hakusanoja ja hakutesauruksen niille osoittamia synonyymeja, suppeampia termejä ja rinnakkaistermejä. Tulokset näyttävät myös yleisesti, mikä merkitys semanttisiin suhteisiin perustuvalla hakujen laajentamisella on sanahaussa. Vertailukohdaksi voi ottaa hakujen laajentamisen tilastollisin perustein, joka ei ole ollut kovin menestyksekästä (ks. esim. Ekmekcioglu ym. 1992; Harman 1992; Peat & Willett 1991). Tosin eri testien tulosten vertailuihin liittyy aina varauksia.

Kun hakujen JA-operaattorit korvattiin läheisyysoperaattoreilla, hakutesauruksesta oli eniten hyötyä: perushaun ja laajimman haun tarkkuudessa ei ollut suurta eroa, mutta laajimman haun suhteellinen saanti oli huomattavasti perushaun saantia

parempi. Tulos vastasi odotuksia, sillä esimerkiksi kolmen käsitteen esiintyminen yhdessä virkkeessä on niin tiukka ehto, että jonkin tuloksen saaminen yleensäkin edellyttää melko runsasta käsitteiden vaihtoehtoisten kuvausten käyttöä haussa. Pelkästään yhtä hakutyyppiä tarkasteltaessa kävi ilmi, että läheisyysoperaattorien käyttö heikentää saantia huomattavasti verrattuna JA-operaattorin käyttöön. Vähennys selittyy tekstin ominaisuuksista käsin: Sanomalehtiartikkeleissa sekä kappaleet että virkkeet ovat yleisesti lyhyitä. Tekstin rakenteeseen perustuvien rajausten asemasta hakusanojen rajaaminen tietyn sanamäärän päähän toisistaan voi säilyttää saannin tason paremmin.

Perushaun laajentaminen rinnakkaistermien avulla oli tämän tutkimuksen mukaan tuloksellisinta saannin kannalta. Tosin erot eri termiryhmien välillä eivät olleet suuret (Kuvio 2). Suppeammat termit eivät lisää tulosjoukkoa paljoakaan, mutta eivät liioin heikennä tarkkuutta. Koska synonyymit tulkittiin väljästi, hakutesauruksessa termien jako synonyymeihin ja rinnakkaistermeihin ei ole yksikäsitteistä, mikä osaltaan vaikeuttaa näiden ryhmien vertailua. Tesaurusten tavanomainen termijako ei välttämättä toimi hyvin hakutesauruksessa. Parempia ratkaisuna voivat olla joko semanttisten suhteiden yksityiskohtaisempi jaottelu tai termien välisten suhteiden ilmaiseminen pelkästään eriasteisina läheisyyksinä (vrt. Fox 1980; Paice 1991).

Hakutesaurus on selvästi saantia parantava väline. Tutkimuksessa kuitenkin tarkkuuden keskimääräinen taso oli laajimmassakin haussa melko korkea (51,2%), vaikka hakukysymyksissä ei keskimäärin ollut kuin kolme leikkaavaa käsitettä. Testissä hakulaajennokset tehtiin mekaanisesti kaikilla hakutesauruksen osoittamilla mahdollisilla termeillä, mikä merkitsee, että niistä kaikki eivät liittyneet hakukysymyksen aiheeseen eli eri käsitteiden yhdistelmästä syntyvään kokonaisuuteen. Jos hakija valikoisi lisättävät termit, karsiutuisivat ainakin tämän tyyppiset väärät termit pois ja tarkkuus saattaisi parantua saannin kärsimättä.

Tekstitietokannat ovat useammin suoraan tiedontarvitsijoiden käytössä kuin viitetietokannat. Satunnaiselle tai aloittelevalle tiedonhakijalle sanahaun kaikkien mahdollisuuksien - etujen ja virhelähteiden - huomioon ottaminen on vaikeaa. Suorasaantinen hakutesaurus palvelisi tällaista käyttäjää. Kuten kaikkien tesaurusten, myös hakutesauruksen kokoaminen on työlästä ja siten kallista - hakujen määrä ja saantitoiveet vaikuttavat työn kannattavuuteen. Hakutesaurus on valikoima kaikista mahdollisista käsitteiden välisistä suhteista ja niitä ilmaisevista sanoista. Tiedontarvitsijoiden käsitys esimerkiksi hyödyllisistä assosiaatiosuhteista saattaa vaihdella suuresti, sillä vaikka hakujen yleisteema olisi yhteinen, ovat hakuaiheet ja relevanssitulkinnat aina yksilöllisiä. Jokaisen hakijan tulisi voida rakentaa oma käsitemallinsa hakutesaurukseksi.

Kiitokset.

Kiitän lämpimästi Aamulehteä ja Sanomalehtien ilmoitustoimistoa saamastani avusta aineiston keruussa.

Hyväksytty julkaistavaksi 22.8.1993.

Lähteet

Allan, K. (1986), Linguistic meaning. Volume 1. London: Routledge & Kegan Paul.

Aitchison, J. & Gilchrist, A. (1987), Thesaurus construction. London: Aslib.

Alkula, R. & Honkela, T. (1992), Tekstin tallennus- ja hakumenetelmien kehittäminen suomen kielen tulkintaohjelmien avulla. FULLTEXT-projektin loppuraportti. Espoo: Valtion Teknillinen Tutkimuskeskus, julkaisuja 765.

Baldinger, K. (1980), Semantic theory. Oxford: Basil Blackwell.

Blair, D.C. & Maron, M.E. (1985), An evaluation of retrieval effectiveness for full-text document-retrieval system. Communications of the ACM 28(3): 289–299.

Ekmekcioglu, F.C., Robertson, A.M. & Willett, P. 1992, Effectiveness of query expansion in ranked-output document retrieval system. Journal of Information Science 18(2): 139–147.

Fox, E.A. (1980), Lexical relations: Enhancing effectiveness of information retrieval system. ACM - SIGIR Forum, 15(3): 5–36.

Fox, E.A. (1988), Improved retrieval using a relational thesaurus for automatic expansion of extended Boolean logic queries. Teoksessa: Martha W. Evens (toim.) Relational models of the lexicon: representing knowledge in semantic networks. Cambridge: Cambridge University Press, ss. 199–210.

Haarala, R. (1981), Sanastotyön opas. Kotimaisten kielten tutkimuskeskuksen julkaisuja 16. Helsinki: Valtion painatuskeskus.

Harman, D., 1992, Relevance feedback revisited. Teoksessa Nicholas Belkin, Peter Ingwersen & Annelise Mark Pejtersen (toim.) Proceedings of the fifteenth annual international ACM SIGIR conference on research and development in information retrieval, Copenhagen, Denmark, June 21.–24. 1992, ss. 1–10.

ISO 2788 (1986), Documentation – Guidelines for the establishment and development of monolingual thesauri. Geneve: International Organization for Standardisation.

Jackson, L. (1983), Searching full-text databases. Teoksessa Proceedings of the 7th International Online Information Meeting: 1983 Dec. 6–8; London, England. Oxford: Learned Information, ss. 419–426.

Karlsson, F. (1980), Johdatusta yleiseen kielitieteeseen. Helsinki: Gaudeamus.

Keen, M.E. (1991), The use of term position devices in ranked output experiments. Journal of Documentation 47(1): 1–22.

Keen, M.E.(1992), Some aspects of proximity searching in text retrieval systems. Journal of Information Science 18(2): 89–98.

Kinnucan, M.T., Nelson, M.J. & Allen, B.L. (1987), Statistical methods in information science. Teoksessa Martha E. Williams (toim.), Annual rewiev of information science and technology, vol 22, Amsterdam: Elsevier science publishers, ss. 147–178.

Kristensen, J. & Järvelin, K. (1990), The effectiveness of a searching thesaurus in free-text searching of a full-text database. Internatonal Classification 17(2):77–84.

Lancaster, F.W. (1968), Information retrieval systems: Characteristics, testing, and evaluation. New York: John Wiley & Sons.

Lancaster, F.W. (1972), Vocabulary control for information retrieval. Washington: Information Resources Press.

Lancaster, F.W. (1986), Vocabulary control for information retrieval. (2nd ed.) Arlington: Information Resources Press.

Lu, Xin (1990). Document retrieval: A structural approach. Information Processing and Management, 26(2), 209–218.

McDonald, M. (1984), BASIS – innovation in data management. Teoksessa C.-C. Chen & P. Hernon (toim.) Numeric databases. Norwood: Ablex, ss.219–236.

Paice, C.D. (1991), A thesaural model of information retrieval. Information Processing & Management, 27(5), 433–447.

Peat, H.J. & Willett, P. 1991, The limitations of term co-occurence data for query expansion in document retrieval systems. Journal of the American Society for Information Science 42(5): 378–383.

Piternick, A.B. (1984), Searching vocabularies: a developing category of online search tools. Online review 8(5): 441–449.

Siegel, S. & Castellan, N.J. (1988), Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Soergel, D. (1985), Organizing information: principles of database and retrieval systems. Orlando: Academic Press.

Sormunen, E. & Alkula, R. (1990), Suomenkielisten tekstitietokantojen tallennus- ja hakutekniikkojen kehittäminen: Esitutkimusraportti. Espoo: Valtion teknillinen tutkimuskeskus, Tiedotteita 1121.

Tague-Sutcliffe, J. (1992), The pragmatics of information retrieval experimentation, revisited. Information Processing and Management 28(4): 467–490.

Tenopir, C. & Ro, J. S. (1990), Full text databases. Westport: Greenwood Press.

Tenopir, C. & Shu, M. E. (1989), Magazines in full text: uses and search strategies. Online Review 13(2): 107–118.

Wang, Y-C, Vandendorpe, J. & Evans, M., (1985), Relational thesauri in information retrieval. Journal of American Society for Information Science 36(1): 15–27.

Painamaton lähde

Kristensen, J. (1992), Vapaasanahakujen laajentaminen hakutesauruksen avulla haettaessa indeksoimattomasta tekstitietokannasta. Kirjastotieteen ja informatiikan lisensiaatintutkimus. Tampere: Kirjastotieteen ja informatiikan laitos.

Viite

1 Substantiiviajattelun ylivaltaa kuvaa mainiosti sekin, että hakutesaurukseen tuli vain vähän adjektiiveja ja verbejä ja nekin esiintyivät siis enimmäkseen synonyymiryhmässä.