Jaana Kristensen

Tesauruksen rooli vapaatekstihaussa

Kristensen, Jaana, Tesauruksen rooli vapaatekstihaussa [The role of thesaurus in free text search]. Kirjastotiede ja informatiikka 8(3): 67—72, 1989.

Such semantic features of natural language as synonyms, antonyms, quasi-synonyms and homonyms cause difficulties in free-text searching. It has been suggested that searching thesauri could be used as search aids to overcome these problems. To test this suggestion a searching thesaurus rich in synonyms and quasi-synonyms was constructed and a set of searches was conducted and analysed in a large Finnish full-text database of newspaper articles. The original queries were elicited from journalists who also evaluated the relevance of the retrieved articles. Each query was searched in three distinct modes. In the first mode, the queries included terms of the original search questions only; in the second, synonyms given by the thesaurus were added to the original terms; in the third, also related terms given by the thesaurus were added. The results of these searches are shown in this paper in terms of recall and precision. By setting the relative recall of the third mode searches at 100 %, the average relative recalls of modes one and two were about 45 % and 74 %, respectively. The average precision values for modes one, two and three were 51 %, 44 % and 35 %, respectively. This indicates that by assigning the synonyms given by the thesaurus to the query improves query effectiveness considerably with a negligible decline in precision. The value of related terms is more controversial.

Address: Kalliokuja 7, SF-78400, Varkaus, Finland.

Johdanto

Tietotekniikan kehittyminen on mahdollistanut yhä suurempien tekstikokonaisuuksien elektronisen tallentamisen. Tekstitietokannan sisältävät dokumenttien osia tai mahdollisesti koko tekstin. Ne suuntautuvat selvemmin tiedontarvitsijoille kuin viitetietokannat, koska ne tarjoavat tietoja — dokumentteja — viitteiden asemasta.

Tiedonhaku tekstitietokannoista perustuu samoihin menetelmiin kuin haku viitetietokannoistakin. Indeksitermien käyttö ja vapaatekstihaku ovat tavalliset vaihtoehdot. Edellisestä on kysymys silloin, kun dokumentin sisältö on kuvailtu kontrolloitua sanastoa käyttäen ja haku kohdistuu näitä termejä sisältävään kuvailutermikenttään. Jälkimmäisestä puhuttaessa tallennettavan tekstin kaikki (merkitsevät) sanat ovat haettavissa. Vapaatekstihaku viitetietokannoista voi kohdistua otsikkojen ja abstraktien sanoihin, tekstitietokannoista dokumenttien koko tekstiin.

Tutkimuksia koko tekstiin kohdistuvasta tiedonhausta ei vielä ole paljon — aihe on uusi, koska toimivia tekstitietokantoja on ollut käytössä melko vähän aikaa. Tiedon tallennuksen ja haun tutkimuksessa on kuitenkin usein vertailtu kontrolloitujen ja vapaiden (vapaatekstihaku) termien käytön vaikutusta hakutuloksiin (esim. Swanson 1960; Cleverdon 1967; Hersey, Foster, Stalder & Carlson 1971; Henzler 1978). Yksimielisyyttä parhaan hakutuloksen takaavasta menetelmästä ei ole, mutta usein on todettu vapaiden ja kontrolloitujen termien täydentävän toisiaan (Fugmann 1982, 144; Tenopir 1985, 160). Vapaiden termien ongelmana on

luonnollisen kielen synonyymien ja kiertoilmausten runsaus. Myös geneeristen hakujen tekeminen vapailla termeillä on vaikeaa. (Soergel 1985, 213—214; Dubois 1987, 247.) Kontrolloidun sanaston haittana on hidas uusiutuminen ja kuvailun tarkkuuden heikkeneminen siirryttäessä dokumentin kielestä indeksointikieleen. Indeksointikustannukset ja sanastojen rakentamis- ja ylläpitokustannukset tekevät osaltaan vapaatekstihaun ja indeksoinnista luopumisen houkuttelevaksi. (Dubois 1987, 248; Perez 1982, 187.)

Tiedonhakua nimenomaan tekstitietokannoista ovat viime aikoina tutkineet sekä Tenopir (1985) että Blair ja Maron (1985). Molemmat tutkimukset on tehty suurista, tuhansia dokumentteja sisältävistä tietokannoista. Tenopir vertaili indeksitermeihin, abstrakteihin, otsikkoihin ja koko tekstiin kohdistuvia hakuja keskenään (1985, 152). Hän totesi haun vapailla termeillä dokumentin koko tekstistä tuovan saanniltaan parhaan, mutta tarkkuudeltaan heikon tuloksen. Mikään hakumenetelmistä ei yksinään taannut täydellistä saantia. (Tenopir 1985, 160.) Blair ja Maron tarkastelivat vapaatekstihakujen tuloksia suhteutettuina tiedontarvitsijain saantitoiveisiin. Heidän mukaansa vapaatekstihaku johtaa saanniltaan heikkoon tulokseen. Syynä tähän he pitivät luonnollisen kielen ilmaisumahdollisuuksien runsautta. (Blair & Maron 1985, 295—298.) Tutkimustulosten erot selittyvät osin erilaisista lähtökohdista ja arviointiperusteista johtuviksi (Ro 1988, 77).

Tekstitietokantojen, elektronisen julkaisemisen ja arkistoinnin yleistyessä tulee myös vapaatekstihaku lisääntymään. Dokumenttien indeksointi ei enää ole välttämätöntä, koska voidaan olettaa dokumenttien sisältävän itse itseään parhaiten kuvaavat sanat. Tiedontarvitsijat tulevat yhä useammin tekemään tiedonhakuja itse, välittäjien asemasta. Tällöin hakija tarvitsee tukea hakusanojen valinnassa, jotta tiedonhaku ei olisi liian vaivalloista eikä tuloksiltaan epävarmaa. Yhtenä mahdollisuutena Lancaster (1972, 149) on esittänyt kontrolloidun sanaston käyttöä ainoastaan tiedon hakuvaiheessa. Ratkaisu tarjoaisi kontrolloidun sanaston etuja ilman indeksointikustannuksia.

Kirjallisuudessa esitellään erilaisia vapaatekstihakua tukevia kontrolloituja sanastoja. Ne vaihtelevat synonyymisanastoista useampia tesauruksia yhdistävään tietokantaan. (Aitchinson & Gilchrist 1987, 119; Lancaster 1986, 172; Piternick 1984, 443—446.) Tesaurus on kontrolloitu sanasto, joka ilmaisee käsitteiden väliset, yhteydestä riippumattomat suhteet (ISO 2788/1986, 2). Tesauruksia on käytetty dokumenttien kuvailussa ja sitten apuna etsittäessä hakutermejä. Tesauruksen tehtävänä on ollut kaventaa luonnollisen kielen ilmaisukirjoa ohjaamalla useasta vaihtoehtoisesta termistä yhteen kuvailutermiin. Hakutesaurus toimii päinvastoin — sen tehtävänä on antaa hakijalle vaihtoehtoja hakutermeiksi vapaatekstihaussa, koska kieltä ei ole kontrolloitu tiedon tallennusvaiheessa. (Soergel 1985, 222—224.)

Luonnollisen kielen aiheuttamat hakuongelmat ovat osittain kielen ominaispiirteistä johtuvia, osittain kaikille kielille yhteisiä. Esimerkiksi Suomen kielelle ominaisia piirteitä, kuten sanan taipumista lukuisissa muodoissa, voidaan hallita tiedonhaussa MORFO-ohjelmiston avulla. Se saattaa tekstin sanat perusmuotoihin, joista käänteistiedosto voidaan rakentaa. Hakijan ei näin ollen tarvitse miettiä hakusanansa kaikkia taivutusmuotoja, vaan perusmuoto riittää. (Nurminen 1986, 62—86; ks. myös Jäppinen & Ylilammi 1986.) Tämän jälkeenkin hakija joutuu miettimään, miten hänen hakemansa asia on eri dokumenteissa ilmaistu. Ongelma ei ole sidoksissa yhden kielen ominaisuuksiin vaan liittyy yleensä luonnollisen kielen semantiikkaan. Vaihtoehtoiset hakutermit eli dokumenttien erilaiset ilmaukset voivat olla synonyymejä, kvasisynonyymejä tai assosiaatiosuhteessa olevia termejä. Näiden suhteiden säätely on kontrolloidun sanaston keskeisiä tehtäviä.

Seuraavassa kuvatun tutkimuksen tarkoituksena oli selvittää, miten hakutesauruksen käyttö vaikuttaa vapaatekstihakujen tuloksiin. Sanomalehtiartikkeleita sisältävästä tekstitietokannasta tehtiin tiedonhakuja, joihin hakutermejä etsittiin tesauruksesta. Erityisesti synonyymien ja rinnakkaistermien merkitys vaihtoehtoisina hakutermeinä pyrittiin saamaan esiin. Vastaavaa kokeellista tutkimusta ei tiettävästi ole aikaisemmin tehty. (Ks. Kristensen 1989.)

Koeasetelma

Tutkimuksen tiedonhaut perustuivat tietokannan käyttäjiltä haastatteluissa saatuihin hakukysymyksiin. Hakutesauruksen käytön vaikutus hakuihin pyrittiin selvittämään tekemällä kolme hakusuunnitelmaa: ensimmäisessä hakutermeinä olivat hakukysymyksen sanat, toisessa lisättiin hakutesauruksen antamat synonyymit, kolmannessa tesauruksen rinnakkaistermeistä sopivat. Termien kombinointitavan määräsi tutkija. Toisessa ja kolmannessa haussa alkuperäisiä termejä täydennettiin niiden synonyymien ja lähitermien yhdisteellä (OR) haun yleisrakennetta muutoin muuttamatta.

Ensimmäinen hakusuunnitelma vastaa tiedonhakijan ilmaisua hakuaiheesta. Se on kysymyksen lyhin esitys, joka voidaan vapain termein tietokannasta hakea. Sen jälkeen täytyy hakuun löytää myös muut mahdolliset ilmaisutavat. Iivosen (1989, 177) mukaan eri indeksoijien valitsemien, samaa dokumenttia kuvaavien termien välillä vallitsee usein assosiaatiosuhde. Voidaan olettaa, että eri kirjoittajien samasta aiheesta käyttämien ilmaisujen välillä vallitsee myös assosiaatio- ja ekvivalenssisuhteita. Toinen ja kolmas hakusuunnitelma lisäsivät nämä vaihtoehdot hakuun.

Kokeessa pyrittiin mallintamaan tilannetta, jossa tiedontarvitsija on tiedonhakijana ja käyttää automaattista välinettä. Kokeessa oletettu tiedontarvitsija antoi ajattelemansa hakutermit ja sai tesauruksen niille tarjoamat vaihtoehdot. Ensimmäinen haku ilmaisi vain lähtökohdan, ei haun lopullista muotoilua. Toisten hakusuunnitelmien avulla haluttiin selvittää, voidaanko hakijan kirjoitusvaivaa ja -aikaa säästää hakutesauruksen avulla eikä niinkään vertailla tesauruksen käyttöä ja käyttämättömyyttä.

Tutkimuksessa käytettiin sanomalehtiartikkeleita sisältävää tekstitietokantaa (Satakunnan Kansa). Tiedonhakujärjestelmänä tietokannassa on BASIS (McDonald 1984, 219—224). Käänteistiedosto on rakennettu sanojen perusmuodoista MORFO-ohjelmistoa käyttäen. Tietokannan koko on merkitsevä tulosten yleistettävyyden kannalta (Blair 1986, 18). Satakunnan Kansan tietokannan tutkimuksessa käytetty osa sisälsi artikkelit ajalta 31. 5. 1987—7. 9. 1988, yhteensä noin 34 000 artikkelia.

Sanomalehtiteksti poikkeaa sanastoltaan ja rakenteeltaan tieteellisestä tekstistä. Tesaurusta tai sanastoa, joka olisi sopinut lehtiartikkeleiden hakuun ei ollut käytettävissä. Tutkimusta varten laadittiin pieni, hakutesaurustyyppinen sanasto talousaiheista. Muodoltaan se oli muuten tavallisen tesauruksen kaltainen mutta kuvailutermejä ei eroteltu viittaustermeistä, koska kaikki termit olivat mahdollisia hakutermejä. Tutkimuksen hakutesauruksen termit olivat neljän tyyppisiä: synonyymeja, laajempia ja suppeampia termejä, rinnakkaistermejä. Termejä oli kaikkiaan 328 ja ne edustivat 180 käsitettä.

Hakukysymykset saatiin haastattelemalla taloustoimittajia. Heille selitettiin kokeen perusasetelma ja pyydettiin kysymyksiä, joita he tekisivät tiedonhakuina tekstitietokannasta. Toimittajia oli viisi ja hakukysymyksiä kerättiin kaikkiaan 30.

Tutkimuksen tekijä laati ja toteutti hakusuunnitelmat. Kysymyksen eri käsitteitä ilmaisevat sanat yhdistettiin jokaisessa haussa ja-operaattorilla. Jos hakuterminä oli sanaliitto, haettiin sitä käyttäen läheisyysoperaattoria, joka etsi sanaliiton sanat yhden kappaleen sisältä. Hakusuunnitelmat laajenivat ensimmäisestä kolmanteen. Toinen haku sisälsi ensimmäisen, kolmas taas toisen ja ensimmäisen tulokset. Laajimman eli kolmannen haun saamat artikkelit tulostettiin. Muiden hakujen tuloksista kirjattiin artikkelien otsikot ja päivämäärät. Artikkelit käytiin läpi ja kysymyksen aihetta käsittelemättömät poistettiin joukosta. Kysymyksen esittänyt toimittaja arvioi jäljelle jääneiden artikkelien relevanssin asteikolla hyödyllinen, hyödytön tai vaikea sanoa. Relevantteja artikkeleita saatiin vastauksena 26 kysymykseen.

Hakutulosten arviointiperusteet

Tiedonhakututkimuksissa on yleisesti käytetty tulosten arvioinnin perusteena saantia ja tarkkuutta. Edellinen ilmaisee kuinka monta prosenttia kaikista tietokannan sisältämistä relevanteista dokumenteista on hakutuloksessa. Todellista saantia on lähes mahdotonta laskea, koska tietokannan kaikkia dokumentteja ei voida tarkastaa. Yleensä käytetäänkin saannin arviota eli suhteellista saantia. Tarkkuus puolestaan ilmaisee kuinka monta prosenttia saaduista dokumenteista on relevantteja. Saannin ja tarkkuuden pätevyyttä tulosten ainoina mittareina on arvosteltu (Saracevic 1975; Swanson

1988), mutta muita vaihtoehtoja ei juuri ole esitetty. Tässäkin tutkimuksessa tulosten arviointi perustui saantiin ja tarkkuuteen.

Koska hakuja vertailtiin keskenään, ei absoluuttista saantia pyrittykään selvittämään. Laajimman eli kolmannen hakusuunnitelman saanti otettiin saannin arvioksi ja sille annettiin arvo 100 %. Ensimmäisen ja toisen haun saantia verrattiin siihen. Tutkimus ei näin ollen kerro, kuinka paljon tietokannan sisältämistä relevanteista artikkeleista saatiin. Pyrkimyksenä oli selvittää, miten tietyntyyppisten hakutermien lisääminen vaikuttaa saantiin ja tarkkuuteen.

Suhteellinen saanti ja tarkkuus laskettiin jokaiselle haulle toimittajien relevanssiarvion mukaan. Asteikko oli kolmiportainen: relevantti, »vaikea sanoa», epärelevantti. Relevanteiksi kelpuutettiin vain relevanteiksi merkityt artikkelit, ei »vaikea sanoa» -tapauksia.

Tulokset

Koska tutkimuksessa käytetty hakutesaurus ei ollut kovin laaja, ei siitä löytynyt kaikkiin kysymyksiin lisätermeiksi sekä synonyymeja että rinnakkaistermejä. Tästä syystä tuloksia tarkastellaan toisaalta koko aineiston osalta (26 kysymystä), toisaalta niiden kysymysten osalta, joita voitiin laajentaa sekä synonyymeillä että rinnakkaistermeillä (18 kysymystä).

Koko aineiston kaikkien saatujen artikkelien määrä vaihteli 1:stä 51:een. Eri hakutyypeissä kaikkien saatujen ja relevanttien artikkelien keskiarvot olivat seuraavat: I haussa (hakukysymyksen sanat) 9,3 ja 4,0, II haussa (edelliset + synonyymit) 19,2 ja 7,9, III haussa (edelliset + rinnakkaistermit) 28,5 ja 9,4.

Saanti ja tarkkuus suhtautuvat toisiinsa yleensä käänteisesti (Lancaster 1986, 133). Myös tässä aineistossa suppeimman haun tarkkuus oli paras (painottamaton keskiarvo 50,6 %) ja saanti heikoin (painottomaton ka 45,2 %). Tarkkuuden keskiarvo putosi synonyymien lisäyksellä 43,6 %:iin ja saanti vastaavasti kasvoi 74,1 %:iin. Rinnakkaistermien lisääminen kasvatti saantia edelleen, mutta tarkkuus heikkeni 34,9 %:iin. (Ks. taulukko 1.)

Taulukko 1. Suhteellisen saannin ja tarkkuuden keskiarvot eri hakutyypeissä.

Keskiarvot	Hakutyypit		
	I[a]	II[b]	III[c]
Suhteellinen saanti %	45,2	74,1	100
Tarkkuus %	50,6	43,6	34,9

[a] Haku I = hakutermeinä kysymyksen sanat
[b] Haku II = hakutermeinä edelliset + tesauruksen synonyymit
[c] Haku III = hakutermeinä edelliset + tesauruksen rinnakkaistermejä

Keskiarvojen eron tilastollisen merkitsevyyden testaamiseen käytettiin t-testiä. Hakujen I ja II saannin keskiarvojen ero oli tilastollisesti erittäin merkitsevä (t-testi suure = 4,05). Tarkkuuden keskiarvojen ero ei hakujen I ja II välillä ollut tilastollisesti merkitsevä (t = —1,37), hakujen I ja II välillä ero oli erittäin merkitsevä (t = —4,43). Haku kysymyksen sanoilla ja synonyymeilla osoittautui kannattavaksi: saanti oli tässä haussa hyvä, lähellä laajimman haun saantia eikä tarkkuuskaan ollut paljon parasta tarkkuuden arvoa heikompi.

Mielenkiintoinen osa-aineisto ovat ne 18 hakukysymystä, joista tehtiin kaikki kolme hakusuunnitelmaa. Koko aineistossa osasta kysymyksiä tehtiin I haun lisäksi vain II tai III haku. Koska osa-aineistossa alkuperäiset kysymykset olivat samat kaikissa hauissa, on tulosten vertailu yksiselitteistä. Näiden 18 kysymyksen osalta suhteellisen saannin keskiarvot ensimmäisessä haussa 44,4 %, toisessa 76,8 % ja kolmannessa 100 %. Tarkkuuden keskiarvot olivat vastaavasti I haussa 48,5 %, II haussa 40,2 % ja III haussa 35,0 %.

Lisäksi haluttiin selvittää, mikä oli eri termityyppien lisäämisen vaikutus tuloksiin. Jokaisesta hausta erotettiin aikaisemmassa haussa saadut tulokset ja jäljelle jäivät kunkin haun lisätulokset. Kaikkien artikkelien määrä oli kutakuinkin yhtä suuri kaikissa hauissa, kun edellisen haun tulokset oli vähennetty. Relevanttien artikkelien määrä sen sijaan vaihteli enemmän. (Ks. Taulukko 2.)

Taulukko 2. Osa-aineiston tulokset: eri hakutermiryhmien saamat lisätulokset.

Haut	Keskiarvot			
	Kaikki art. lkm.	Relevantit art. lkm.	Suhteellinen saanti %	Tarkkuus %
I[a]	9,2	4,1	44,4	48,5
II—[b]	9,5	3,5	32,4	30,7
III—[c]	9,7	1,5	23,2	12,9

[a] Haku I = hakutermeinä kysymyksen sanat
[b] Haku II— = hakutermeinä edelliset + tesauruksen synonyymit, tuloksista poistettu ensimmäisessä haussa saadut artikkelit
[c] Haku III— = hakutermeinä edelliset + tesauruksen rinnakkaistermejä, tuloksista poistettu kahdessa ensimmäisessä haussa saadut artikkelit.

Näin tarkasteltuina suhteellinen saanti ja tarkkuus ovat parhaat haettaessa kysymyksen sanoilla ja heikoimmat rinnakkaistermihaussa. Ero hakujen I ja III- suhteellisen saannin keskiarvojen välillä on tilastollisesti melkein merkitsevä ($t = -2{,}73$). Hakujen I ja II- saannin ero ei ole tilastollisesti merkitsevä ($t = -1{,}70$). Tarkkuuden suhteen hakujen I ja III- keskiarvojen ero on tilastollisesti erittäin merkitsevä ($t = -8{,}63$), hakujen I ja II- välillä ero on melkein merkitsevä ($t = -2{,}54$). Toinen haku on kokonaistuloksen kannalta tärkeämpi kuin kolmas: toisessa haussa saadaan enemmän uusia relevantteja artikkeleita kuin kolmannessa ja epärelevanttien osuus on pienempi.

Yhteenveto

Tutkimuksessa tarkasteltiin tesauruksen esittämien synonyymien ja rinnakkaistermien liittämistä vapaatekstihakuun vaihtoehtoisina hakutermeinä. Kolmentyyppisiä hakuja verrattiin toisiinsa: 1, hakuja, jotka perustuivat hakukysymysten sanoihin, 2, hakuja, joihin edellisten hakutermien lisäksi liitettiin hakutesauruksen synonyymit, 3, hakuja, joihin edellisten termien ohella lisättiin hakutesauruksen rinnakkaistermejä. Tärkein havainto oli, että hakutesauruksen käyttö parantaa merkittävästi hakutulosten saantia. Synonyymien liittäminen hakuun oli kannattavaa, koska saanti parani tuntuvasti (45,2 %:sta 74,1 %:iin) eikä tarkkuus heikentynyt ratkaisevasti (50,6 %:sta 43,6 %:iin). Rinnakkaistermien lisääminen sen sijaan heikensi tarkkuutta huomattavasti, joskin relevantteja dokumentteja saatiin vielä lisää.

Hakutesauruksen avulla pyritään ratkaisemaan luonnollisen kielen semanttisen tason ongelmia. Se auttaa hakijaa löytämään merkitykseltään samankaltaisia tai läheisiä sanoja hakutermeiksi. Näitä etuja ei voida saavuttaa sanoja perusmuotoihin palauttavien tai taivutusmuotoja luovien ohjelmien avulla. Kun tällaiset ohjelmat perustuvat kieliopin (syntaksi) sääntöihin, perustuu hakutesaurus käsitteelliseen mallintamaiseen (semantiikka). Tästä johtuu myös se, että hakutesauruksen käyttökelpoisuutta koskevat tutkimustulokset eivät ole kielisidonnaisia kuten syntaksia soveltavien ohjelmien käyttökelpoisuutta koskevat tulokset.

Tiedonhaun välineiden kehittäminen on ajankohtaista tiedon tallennuksen ja haun tutkimuksessa. Hakutesaurus vaatii edelleen kehittämistä. Tämän tutkimuksen ulkopuolelle jäivät hierarkisten suhteiden merkitys vapaatekstihaussa samoin kuin läheisyysoperaattorin käyttö tarkkuuden parantamiseksi. Hakutesauruksen käyttökelpoisuus on kuitenkin tullut tutkimuksessa osoitetuksi.

Kiitokset. Tutkimustyötäni ovat auttaneet apulaisprofessori Kalervo Järvelin, lehtori Mirja Iivonen ja Markku Ylinen monilla neuvoillaan ja tekstin parannusehdotuksillaan. Olen saanut tukea myös C. V. Åkerlundin säätiöltä.

Hyväksytty julkaistavaksi 11. 9. 1989.

Lähteet

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

Blair, D. C.: Full text retrieval: evaluation and implications. International Classification 13(1) 1986, 18—23.

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

Cleverdon, C.: The Granfield tests on index language devices. Aslib Proceedings 19(6) 1967, 173—194.

Dubois, C. P. R.: Free text vs. controlled vocabulary; a reassessment. Online Review 11(4) 1986, 243—253.

Fugmann, R.: The complementarity of natural and indexing languages. International Classification 9(3) 1982, 140—144.

Henzler, R. G.: Free or controlled vocabularies; some statistical user-oriented evaluations of biomedical information systems. International Classification 5(1) 1978, 21—26.

Hersey, D. F., Foster, W. R., Stalder, E. W. & Carlson, W. T.: Free text word retrieval and scientist indexing: performance profiles and costs. Journal of Documentation 27(3) 1971, 167—183.

Iivonen, M.: Indeksointituloksen riippuvuus indeksointiympäristöstä. Tampereen yliopiston Kirjastotieteen ja informatiikan laitoksen tutkimuksia, 26. Tampere, 1989.

ISO 2788. Documentation — Guidelines for the establishment and development of monolingual thesauri. International Organisation for Standardisation, 1986.

Jäppinen, H. & Ylilammi, M.: Associative model of morphological analysis: an empirical inquiry. Computational Linquistics 12(4) 1986, 257—272.

Kristensen, J.: Tesauruksen rooli vapaatekstihaussa. Kirjastotieteen ja informatiikan pro gradu -tutkielma. Tampere, 1989.

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

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

McDonald, M.: BASIS-innovation in data management. Teoksessa C. Chen ja P. Hernon (toim.) Numeric Databases. Norwood: Ablex, 1984.

Nurminen, R.: Suomen kielen sanamuotoja tulkitsevien ohjelmien hyödyntäminen tiedonhakujärjestelmissä. Tutkimuksia 386. Espoo: Valtion teknillinen tutkimuskeskus, 1986.

Perez, E.: Text enhancement: controlled vocabulary vs. free text. Special Libraries 73(3) 1982, 182—192.

Piternick, A.: Searching vocabularies: a developing category of online searching tools. Online Review 8(5) 1984, 441—449.

Ro, J. S.: An evaluation of the applicability of ranking alcorithms to improve the effectiveness of full-text retrieval; 1. On the effectiveness of full-text retrieval. Journal of the American Society for Information Science 39(2) 1988, 73—78.

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

Soergel, D.: Organizing information: principles of database and retrieval systems. New York: Academic Press, 1985.

Swanson, D. R.: Seraching natural language by computer. Science 132(3434) 1960, 1099—1104.

Swanson, D. R.: Historical note: Information retrieval and future of an illusion. Journal of the American Society for Information Science 39(4) 1988, 92—98.

Tenopir, C.: Full-text database retrieval performance. Online Review 9(2) 1985, 149—164.