Googleta se!

iGS-verkkotietopalvelun kysymystyypit ja vastausten löytyminen hakukoneiden keskivertokäyttäjien ”simulaation” avulla

Jani Keränen

Tampereen yliopisto
Informaatiotutkimuksen ja
interaktiivisen median laitos
Pro gradu -tutkielma
Huhtikuu 2010

TIIVISTELMÄ

TAMPEREEN YLIOPISTO
Informaatiotutkimuksen ja interaktiivisen median laitos
JANI KERÄNEN: Googleta se! iGS-verkkotietopalvelun kysymystyypit ja vastausten löytyminen hakukoneiden keskivertokäyttäjien ”simulaation” avulla.
Pro gradu –tutkielma, 66 s.
Informaatiotutkimus
Huhtikuu 2010

Tutkielmassa selvitettiin, missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS-verkkotietopalvelun kysymyksiin, ja minkätyyppisiä kysymyksiä asiakkaat lähettävät iGS-palveluun. Lisäksi tutkimuksessa selvitettiin, onko iGS:n kysymysten tyypillä yhteyttä siihen että missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin.

Tutkimusmetodina käytettiin tätä tutkimusta varten kehitettyä hakukoneiden keskivertokäyttäjien ”simulaatiota”. Hakukoneiden käyttäjistä saatua tutkimustietoa sovellettiin eräänlaisten tutkimuksen ennakkoehtojen ja rajausten luomisessa. Käyttäjätutkimusten tulosten perusteella muodostettiin hakukoneiden keskivertokäyttäjiä ”simuloivia” sääntöjä, joiden tarkoituksena oli vähentää hakijan yksilöllisten ominaisuuksien vaikutusta hakutuloksiin. ”Simulaation” säännöissä määriteltiin mm. kyselyissä käytettävien hakutermien valinta, määrä ja muoto, edistyneempien hakutekniikoiden käyttö, kyselyiden muodostamismenetelmä, hakujen määrä sekä hakutulosten läpikäyminen.

Hakukoneiden keskivertokäyttäjien ”simulaation” avulla Google-hakukoneella löytyi relevantti suomenkielinen vastaus 188 kysymykseen. Tutkimuksen otoksen 509 alakysymyksestä tämä oli 36,9 prosenttia, eli runsas kolmasosa. ”Simulaation” säännöissä määriteltyjen rajoitusten valossa tämä tuntuu varsin suurelta luvulta. 321 kysymykseen (63,1 prosenttia) ei vastausta löytynyt Googlella.

Otoksen kysymykset jaettiin kolmeen kysymyskategoriaan. Tiettyä aihetta koskevia kysymyksiä oli eniten (44,2 prosenttia), mutta lähes yhtä paljon oli faktakysymyksiä (43,6 prosenttia). Tiettyä teosta koskevia kysymyksiä oli selvästi vähiten (12,2 prosenttia). Faktakysymyksiin löytyi parhaiten vastaus Googlella (41,4 prosenttia), seuraavaksi parhaiten tiettyä aihetta koskeviin kysymyksiin (33,8 prosenttia), ja huonoiten tiettyä teosta koskeviin kysymyksiin (30,5 prosenttia). iGS:n kysymysten tyypeillä ei kuitenkaan ollut tilastollisesti merkittävää yhteyttä siihen että missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin.

Hakukoneiden keskivertokäyttäjien ”simulaatiosta” saatuja tuloksia on melko vaikeaa yleistää, koska tällaista tutkimusmetodia ei ole koskaan aikaisemmin käytetty. Metodilla saatiin kuitenkin aikaiseksi suuntaa-antava ”vähimmäistulos”, ja lisäksi saatiin tietoa palvelun kysymystyypeistä. Eräänlaisena tutkimustuloksena voidaan pitää myös uudenlaisen tutkimusmetodin aikaan saamista. Hakukoneiden keskivertokäyttäjien ”simulaatio” on puutteineenkin melko käyttökelpoinen tutkimusmetodi mahdollisten jatkotutkimusten kannalta. Sitä voidaan myös kehittää edelleen, ja soveltaa tiedonhakuun liittyvissä tutkimuksissa.

Avainsanat (YSA): yleiset kirjastot, verkkotietopalvelu, kysymykset, luokitukset, tiedonhaku, hakuohjelmat, Google, käyttäjät, käyttäjätutkimus, simulointi

SISÄLLYS

1. JOHDANTO

Internet-yhteyksien ja varsinkin laajakaistayhteyksien yleistymisen myötä internetin käyttö on lisääntynyt Suomessa valtavasti viime vuosien aikana. Erityisesti arkielämän tiedonhankinnassa internetin merkitys on kasvanut lähes eksponentiaalisesti. Tiedonhaku netistä on yhä useammalle jo arkipäivää.

Internetissä yleisimmin tiedonhakuun käytettyjä välineitä ovat sanahakupalvelut eli hakukoneet. Nykyisin ylivoimaisesti suosituin hakukone on Google. Yli 70 prosenttia kaikista internet-hauista tehdään Googlen avulla Lazulyn (2007, 106) mukaan. Peräti 95 prosenttia suomalaisista käytti säännöllisesti hakukoneenaan Googlea vuonna 2007. Googlen suosiota kuvaa hyvin se, että varsin monelle tiedonhaku netistä tarkoittaa samaa asiaa kuin ”googlettaminen”.

Toisinaan internetin hakukoneet on nähty uhkana perinteisille tiedonhankintakanaville, kuten esimerkiksi kirjastolle ja sen tietopalvelulle. Koska hakukoneilla voidaan tyydyttää erilaisia tiedontarpeita helposti ja nopeasti jo kotoa käsin, periaatteessa ei ole syytä ”lähteä merta edemmäs kalaan”, eli kääntyä esimerkiksi kirjaston tietopalvelun puoleen. Vaikka Suomessa ei vielä ole nähtävissä selviä merkkejä verkkotietopalvelun suosion vähenemisestä, tilanne voi kuitenkin muuttua lähitulevaisuudessa. Internetistä löytyvän tiedon määrä kasvaa päivä päivältä, ja hakukoneet kehittyvät yhä tehokkaammiksi ja helppokäyttöisimmiksi. Ne voidaan jo nyt nähdä vakavasti otettavana kilpailijana verkkotietopalvelulle – mitä ne ovat sitten tulevaisuudessa?

Hyvin todennäköisesti internetin hakukoneet eivät tule koskaan täysin syrjäyttämään kirjastoa ja sen tietopalvelua. Kaikki eivät osaa tai halua käyttää internetin hakukoneita, vaan mieluummin kääntyvät tiedonhaun ammattilaisen puoleen. Kaikkea tietoa ei myöskään löydä netistä, oli hakija kuinka hyvä tahansa. Kirjaston tietopalvelua ja verkkotietopalvelua tullaan aina luultavasti tarvitsemaan. Hakukoneiden kirjastojen palveluille muodostamaa ”uhkaa” ei kuitenkaan kannata vähätellä. Päinvastoin olisi ehkä syytä pohtia, kuinka tarpeellisia ja tehokkaita nykyiset verkkotietopalvelut ovat ylipäänsä? Nykyisiä kirjaston verkkotietopalveluja ja niiden käyttöä ei ole vielä tutkittu kovinkaan paljoa, vaikka palvelujen tuottaminen ja ylläpitäminen vaativat paljon resursseja. Koska internetin hakukoneet voidaan nähdä kirjaston verkkotietopalvelun kilpailijoina, olisi syytä ainakin selvittää, kuinka hyvin nykyiset hakukoneet – esimerkiksi Google – pärjäävät

kirjaston verkkotietopalvelulle. Löytyykö Googlella todellakin helposti vastaus kaikkiin ongelmiin – myös niihin kysymyksiin, joita lähetetään kirjaston verkkotietopalveluun?

Tämän tutkimuksen päätavoitteena on selvittää, että missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia kirjaston verkkotietopalvelun kysymyksiin. Lisäksi tutkimuksessa selvitetään, minkätyyppisiä kysymyksiä asiakkaat lähettävät verkkotietopalveluun, ja onko kysymysten tyypillä yhteyttä Googlella saatuihin hakutuloksiin.

Tutkimuksen kohteeksi on valittu iGS Tietohuoltoasema (information Gas Station) ja sen julkinen kysymys-vastaus –arkisto. iGS on Helsingin kaupunginkirjaston maksuton verkkotietopalvelu. Sen tunnuslause on ”Kysy mitä vain”. iGS-palvelusta voi todella kysyä aivan mitä tahansa. Kysymysten ja vastausten monipuolisuuden vuoksi iGS:n kysymys-vastaus –arkisto tarjoaakin mielenkiintoisen tutkimuskohteen. Palveluun lähetetään kaikenlaisia kysymyksiä – aivan kuten hakukoneidenkin avulla etsitään vastauksia kaikenlaisiin ongelmiin. iGS-palvelu soveltuu myös toisesta syystä hyvin internetin hakukoneiden tutkimiseen: monissa iGS:n vastauksissa on käytetty lähteinä internetistä löytyviä sivustoja, jotka ovat periaatteessa löydettävissä hakukoneilla.

Tutkielma rakentuu seuraavasti: Tärkeimpien käsitteiden määrittelyn jälkeen luvussa 2 tarkastellaan verkkotietopalvelujen kysymysten luokitteluvaihtoehtoja, ja käydään läpi yleisten kirjastojen verkkotietopalveluja tarkastelevia tutkimuksia. Lisäksi tässä luvussa perehdytään hakukoneiden käyttäjätutkimuksiin. Kolmannessa luvussa esitellään empiirisen tutkimuksen asetelma. Lisäksi tässä luvussa perehdytään tarkemmin tutkimuksen kohteeseen eli iGS Tietohuoltoasemaan sekä tutkimuksen välineeseen eli Googleen. Neljännessä luvussa raportoidaan tutkimuksen tulokset. Tulosten yhteenveto ja johtopäätökset esitetään luvussa 5.

2. TEOREETTINEN VIITEKEHYS

2.1. Käsitteiden määrittelyä

Tietopalvelu

Sanastokeskus TSK määrittelee tietopalvelun seuraavasti: Tietopalvelu on tietohuollon osa, joka välittää tietoa sen tarvitsijoille sekä avustaa tiedonlähteiden käytössä. Tietopalvelussa käytetään tiedonlähteinä esimerkiksi kirjastoja, arkistoja ja asiantuntijoita. (Sanastokeskus TSK ry:n verkkosivut, TEPA-termipankki.)

Verkkotietopalvelu

Tietopalvelua on mahdollista saada kahdessa tilassa: yhtäältä kirjaston fyysisissä tiloissa eli kirjastotiskin äärellä tai sen kautta. Toisaalta tietopalvelua tarjotaan ”verkkokirjastossa”, kirjaston tarjoamina verkkopalveluina. Verkkotietopalvelu voidaan määritellä verkon kautta tarjottavaksi ihmiselle välitetyksi avuksi (Lankes, Gross & McClure (2003, 402; tässä Hälinen 2004, 3).

Synkroniset ja asynkroniset palvelut

Neuvonta- ja tietopalvelut voidaan jakaa synkronisiin ja asynkronisiin palveluihin. Molemmat termit juontuvat kreikasta ja tarkoittavat ”samaan aikaan” ja ”ei samaan aikaan”. Synkroninen asiakaspalvelu tapahtuu tässä ja nyt, ilman varsinaista odotus- tai toimitusaikaa: vastauksen saa heti kun kysymyksen on esittänyt. Synkronisia palveluja ovat mm. tiskillä tai puhelimessa tapahtuva neuvonta sekä esimerkiksi chat-palvelu. Asynkronisessa palvelussa asiakkaan toimeksiantoon ei reagoida välittömästi vaan tietyn viiveen tai toimitusajan jälkeen: vastauksen kysymykseen saa myöhemmin, esimerkiksi tuntien tai päivien kuluttua. Helsingin kaupunginkirjastossa asynkronisia palveluja ovat Kysy kirjastonhoitajalta sekä iGS Tietohuoltoasema. (Hälinen 2004, 3; Juntumaa 2003, 3.)

Hakupyyntö

Hakupyynnöllä tarkoitetaan asiakkaan välittäjälle esittämää asiakkaan tiedontarpeen kuvausta, jonka asiakas esittää halutessaan tiedontarpeeseensa liittyvää informaatiota ja jonka välittäjä tiedonhaussa käsittelee edelleen. Asiakkaan on usein vaikeaa ilmaista tarkasti todellinen tiedontarpeensa, joten hakupyyntö ei ole välttämättä sama kuin asiakkaan todellinen tiedontarve. (Iivonen 1995, 7.)

Kysely, hakulauseke

Kysely on hakijan mielessä olevan tiedontarpeen muotoilu sellaiseen muotoon, että järjestelmä ymmärtää sen. Tiedonhaun kirjallisuudessa kysely -nimitystä vastaten esiintyvät usein myös nimitykset hakukysely ja hakulauseke. Hakulauseke on hakupyynnön kuvaus, joka koostuu sekä hakukäsitteitä kuvaavista hakutermeistä että tarvittaessa niitä yhdistävistä operaattoreista. Hakulauseke voi koostua vain yhdestä hakutermistä, mutta myös useista hakutermeistä, jotka on yhdistetty toisiinsa erilaisilla operaattoreilla. *Hakulausekkeiden muotoilulla* tarkoitetaan päätöstä siitä, mitkä käsitteet valitaan hakukäsitteiksi, millä hakutermeillä hakukäsitteet kuvaillaan ja miten nämä hakukäsitteitä edustavat hakutermit yhdistetään toisiinsa. (Alaterä & Halttunen 2002, 34–35; Iivonen 1995, 9–10.)

Hakukäsite

Hakukäsitteillä tarkoitetaan hakulausekkeissa ilmaistuja tiedonyksiköitä, jotka ovat kielen ulkoisen maailman olioita, tapahtumia tai ominaisuuksia edustavia merkityksiä ja joita kuvataan hakutermeillä. Yksi hakukäsite voidaan kuvata useammalla kuin yhdellä hakutermillä. (Iivonen 1995, 8.)

Hakutermi

Hakutermillä tarkoitetaan hakulausekkeessa esiintyvää merkkijonoa. Hakutermi voi sisältää myös katkaisumerkin ja se voidaan erottaa hakulausekkeessa esiintyvistä muista merkkijonoista operaattorilla tai/ja joka aloittaa tai päättää hakulausekkeen. Hakutermistä voidaan käyttää myös nimityksiä *hakuavain* ja *hakusana*. Hakutermi viittaa hakukäsitteeseen. (Iivonen 1995, 9.)

Käsite-, ilmaisu- ja merkkijonotaso

Järvelin (1995, 68–73) on esittänyt, että hakutehtävät ja dokumentit voidaan esittää kolmella tasolla. *Käsitetasolla* tarkastellaan hakutehtävän ja dokumentin käsitteitä ja näiden välisiä suhteita. *Ilmaisutasolla* tarkastellaan käsitteiden ilmaisutapoja luonnollisessa kielessä tai jossakin keinotekoisessa kielessä (kuten dokumentaatiokielessä). Ilmaisutasolla korostuu vaihtoehtoisten ilmausten ideointi ja etsittävien dokumenttien luonne. Hakutehtävän käsitteiden esityksiä ilmaisutasolla kutsutaan *hakuavaimiksi*. Tietokoneisiin perustuva konkreettinen tiedonhaku tapahtuu aina *merkkijonotasolla*. Tietokoneet käsittelevät vain merkkijonoja, joten haun valmistelun viimeinen ja välttämätön vaihe on kyselyn muotoileminen merkkijonotasolla. Hakutehtävä esitetään

tällä tasolla kyselynä, joka koostuu komennoista, operaattoreista, kenttätunnisteista, katkaisumerkeistä jne. Kysely esitetään siis kyseisen tiedonhakujärjestelmän kyselykielellä.

Hakustrategia

Ryhtyessämme etsimään verkosta haluamaamme tiedonlähdettä on ensin luotava hakustrategia. Hakustrategia –termiä on käytetty monessa merkityksessä. Toiset liittävät termin hakulausekkeiden muotoilemiseen ja hakuohjelmalle annettujen komentojen eli kyselyiden tekemiseen. Toiset tarkoittavat hakustrategialla haun kuluessa tehtäviä päätöksiä siitä, mikä on paras tapa jatkaa hakua eli hakutaktiikkaa. Järvelin (1995, 159) määrittelee hakustrategian kokonaissuunnitelmaksi tai lähestymistavaksi haun suorittamiseen, ja hakutaktiikan askeleiksi, jotka suoritetaan valitun strategian edistämiseksi. Järvelinin mukaan yhdellä haulla voi olla vain yksi strategia, mutta siinä voidaan käyttää monia taktiikoita.

Tiedonhakijoiden toimintaa analysoimalla on tunnistettu useita hakustrategiatyyppejä. Hakustrategiatyyppejä ovat mm. pikahaku, lohkohaku, helmenkasvatushaku, lohkojen peräkkäishaku ja selailuhaku. Tässä tutkimuksessa käytetään yksinkertaisinta ja suoraviivaisinta hakustrategiaa eli *pikahakua*. Pikahaussa minimoidaan haun valmisteluaika, ja tyypillisesti siinä käytetään 1–3 hakutermiä. Pikahaku sopii käytettäväksi, kun halutaan löytää vain muutama viite, tunnettu dokumentti tai faktatieto. (Alaterä & Halttunen 2002, 86–87.)

Hakutaktiikka

Hakutaktiikka on tiedonhaun aikana toteutettu yksi tai useampi siirto tiedonhaun jatkamiseksi eteenpäin. Tiedonhaun aikana toteutettu siirto on tunnistettavissa oleva ajatus tai toiminta, joka on osa tiedonhakua. Hakutaktiikkaa tarvitaan kun kysely tuottaa dokumentteja joko liikaa tai liian vähän. Tällöin kyselyä pitää joko kaventaa tai laajentaa. Hakutaktiikoihin liittyy mm. hakutermien etsiminen eri tavoin, hakeminen hierarkkisesti laajemmilla tai suppeammilla hakutermeillä, kirjoitusmuotojen huomioiminen, hakukäsitteiden lisääminen tai vähentäminen, synonyymien ja rinnakkaisten termien käyttö sekä hakutermien rajaaminen kokonaan pois. (Alaterä & Halttunen 2002, 88–89; Järvelin 1995, 226.)

2.2 Aikaisemmat tutkimukset

Tässä luvussa tarkastellaan verkkotietopalvelujen kysymysten luokitteluvaihtoehtoja, sekä teoreettisella että käytännön tasolla. Luvussa käydään läpi yleisten kirjastojen verkkotietopalveluja tarkastelevia tutkimuksia. Erityisesti tarkastelun kohteena ovat tutkimuksissa käytetyt kysymysten taksonomiat. Yhtenä tavoitteena on tunnistaa tämän tutkimuksen kannalta olennaiset kysymysten luokittelukategoriat. Lisäksi 2.2 luvun lopussa perehdytään hakukoneiden käyttäjätutkimuksiin, jotta saataisiin tietoa hakukoneiden ”keskivertokäyttäjistä” tutkimuksen menetelmää eli ”simulaatiota” varten.

2.2.1 Pomerantzin kysymystaksonomiat

Jeffrey Pomerantz (2005) on tarkastellut tunnettuja kysymystaksonomioita tarkastelemalla aihealueen tutkimuskirjallisuutta. Hän keskittyi kysymys-vastaus –järjestelmiin (Question Answering), perinteiseen tietopalveluun ja verkkotietopalveluun sekä lingvistiikkaan liittyviin kysymystaksonomioihin. Kukin näistä alueista käsittelee kysymyksiä eri syistä. Lisäksi kysymyksiä tarkastellaan näillä alueilla melko erilaisina kokonaisuuksina. (Pomerantz 2005, 715.)

Kysymys-vastaus –järjestelmien alueella kysymykset nähdään sellaisina joihin on olemassa ainoastaan yksi oikea vastaus, ja joka on löydettävissä. Vastaus ei välttämättä löydy annetulla järjestelmällä, tai vastaus ei edes sijaitse järjestelmän saatavilla olevassa dokumenttikokoelmassa, mutta vastaus on kuitenkin aina olemassa. Mikäli järjestelmä on hyvä eikä dokumenttikokoelmassa ole puutteita, vastauskin löytyy. (Pomerantz 2005, 715.)

Kysymys-vastaus –järjestelmissä oletetaan, että käyttäjät haluavat suoran vastauksen kysymykseensä (dokumenttien listauksen sijaan). Lisäksi järjestelmissä oletetaan, että vastaus voi löytyä suoraan jostain yksittäisestä tekstinpätkästä, tai monen tekstinpätkän yhdistelmästä. Tästä syystä kysymys-vastaus –järjestelmien tavoitteena on hakea pieniä tekstinpätkiä jotka sisältävät varsinaisen vastauksen. (Pomerantz 2005, 715.)

Tietopalvelussa ja verkkotietopalvelussa kysymystä tarkastellaan täysin eri näkökulmasta. Kysymys nähdään esimerkiksi tiedontarpeena, joka ilmaantuu kun havaitaan aukko yksilön tiedontilassa.

Yksilö ei kuitenkaan välttämättä pysty ilmaisemaan senhetkistä tiedontarvettaan täsmällisesti, eli esitetty kysymys ei välttämättä olekaan se oikea kysymys. Vastaajan tehtävänä on tunnistaa kysyjän varsinainen tiedontarve (esimerkiksi jatkokysymyksien avulla), ja löytää vastaus varsinaiseen kysymykseen. (Pomerantz 2005, 716.)

Myös vastaus nähdään (verkko)tietopalvelussa kysymys-vastaus –järjestelmiä moniulotteisemmin. Hyväksyttävä vastaus riippuu kysyjästä ja hänen senhetkisestä tilanteestaan. Hyödyllinen vastaus auttaa kysyjää selviytymään tilanteesta, josta syntyi kyseinen tiedontarve. Eri yksilöt voivat esittää samanlaisen tai jopa identtisen kysymyksen, mutta vastauksen sisällön ja muodon hyödyllisyys vaihtelee yksilöiden ja tilanteiden mukaan. (Pomerantz 2005, 716.)

Lingvistiikassa kysymykset nähdään kielellisinä ilmiöinä. Niitä voidaan tarkastella Pomerantzin mukaan seitsemällä eri tasolla: fonologinen, morfologinen, leksikaalinen, syntaktinen, semanttinen, diskursiivinen ja pragmaattinen. Kolmella ensimmäisellä tasolla tarkastellaan yksittäisiä sanoja tai niitä pienempiä yksiköitä. Neljällä viimeisellä tasolla tarkastellaan lauseita ja suurempia tekstikokonaisuuksia. Kysymyksiä on mahdollista analysoida ainoastaan näillä neljällä tasolla. Syntaktisella ja semanttisella tasolla tarkastelun kohteena ovat yksittäiset lauseet tai lausahdukset. Diskursiivisella ja pragmaattisella tasolla kieltä tarkastellaan myös ihmisten välisenä vuorovaikutuksena ja viestintänä. (Pomerantz 2005, 717; 723–724).

Pomerantz tunnisti viisi kysymystaksonomiaa, joiden mukaan verkkotietopalveluihin tulleita kysymyksiä voidaan luokitella:

1. Kysymyssanan mukainen luokittelu (Wh-words),
2. Aiheenmukainen luokittelu (Subjects of questions),
3. Kysymyksiin odotettujen vastausten tarkoituksen mukainen luokittelu (The functions of expected answers to questions),
4. Kysymyksiin odotettujen vastausten muodon mukainen luokittelu (The forms of expected answers to questions) sekä
5. Vastausten lähteiden mukainen luokittelu (Types of sources from which answers may be drawn). (Pomerantz 2005, 718.)

Taksonomioista kaksi ensimmäistä perustuu pelkästään kysymysten luokitteluun. Kolmannessa tarkastellaan kysymyksen lisäksi myös vastausta. Kaksi viimeksi mainittua taksonomiaa perustuu vastausten luokitteluun. (Pomerantz 2005: 718–719.)

Kysymyssanan mukainen luokittelu perustuu kysymyksissä esiintyviin englanninkielisiin kysymyssanoihin who, which, what, when, where, why ja how. Tähän luokittelutapaan liittyy kaksi ongelmaa. Ensinnäkin kysymyksiä ei välttämättä ilmaista juuri edellä mainittuja sanoja käyttämällä. Toiseksi toteamukset, joissa käytetään edellä mainittuja sanoja, eivät välttämättä ole kysymyksiä. Kysymyssanoihin perustuva luokittelutapa ei myöskään kerro mitään siitä että mihin kysymys liittyy tai mitä asiakas haluaa. (Pomerantz 2005, 718.) Lisäksi tätä luokittelutapaa on hankalaa (tai jopa mahdotonta) käyttää suomenkielisten kysymysten luokitteluun.

Aiheenmukainen luokittelu on erittäin suosittu tapa luokitella tietopalvelukysymyksiä. Nämä taksonomiat voidaan luokitella kahteen ryhmään. Toisissa käytetään yleistä dokumenttien luokittelujärjestelmää (esim. Deweyn järjestelmä), toisissa taas palvelun järjestäjät ovat itse kehitelleet oman aiheenmukaisen luokittelujärjestelmän (esim. iGS:n luokitusjärjestelmä). (Pomerantz 2005, 719.)

Kolmannessa Pomerantzin tunnistamassa kysymystaksonomiassa, eli kysymyksiin odotettujen vastausten tarkoituksen mukaisessa luokittelussa tarkastellaan sekä kysymystä että siihen liittyvää odotettua vastausta, ennen kuin se itse asiassa on edes muotoiltu. Pomerantzin (2005, 719) mukaan vastaukset usein osoittautuvat monimutkaisemmiksi kuin miltä edeltä käsin vaikuttaa. Hermjakobin (tässä Pomerantz 2005, 719) mukaan kysymyksen luokittelussa on tärkeää ottaa huomioon niiden vastaustyypit. Mikäli kysytään esimerkiksi Mount Everestin korkeutta, ei kysymykseen voi vastata ennen kuin ymmärretään että halutun vastauksen täytyy muodostua mitallisesta määrästä.

Neljännessä taksonomiassa, eli vastausten muodon mukaisessa luokittelussa kysymyksiä tarkastellaan vastaajan näkökulmasta. Kysymyksiä luokitellaan sen mukaan, millaisessa muodossa vastauksen odotetaan olevan, ja kuinka laajaa vastausta kysymys edellyttää. Joissain kysymyksissä odotetaan esimerkiksi pelkästään jonkin teoksen saatavuustietoja, toisissa taas vastaus vaatii laajempaa tutkimustyötä. (Pomerantz 2005, 720.)

Viimeisessä Pomerantzin tunnistamassa taksonomiassa kysymykset luokitellaan vastausten lähteiden mukaan, eli sen mukaan millaisista lähteistä vastaus kysymykseen voidaan löytää. Tähän liittyy vastaajan eli tietopalvelutyöntekijän kyky tunnistaa erilaiset tiedonlähteet, joista vastaus kysymykseen on mahdollista saada. Tähän taksonomiaan liittyy perustavanlaatuinen ongelma: yleensä taksonomioissa pyritään löytämään toisensa poissulkevia kategorioita. Tässä tapauksessa se ei kuitenkaan onnistu, sillä tietopalvelukysymysten vastaukset pohjautuvat usein moniin eri lähteisiin. Pomerantzin mukaan tätä taksonomiaa ei ole varsinaisesti käytetty kysymysten luokittelussa, ehkä juuri edellä mainitusta syystä. (Pomerantz 2005, 721–722.)

Pomerantzin mukaan ainoastaan kolmatta taksonomiaa, eli kysymyksiin odotettujen vastausten tarkoituksen mukaista luokittelua, on käytetty sekä automatisoiduissa kysymys-vastaus –järjestelmissä että kirjastojen tietopalvelututkimuksissa. Monissa automatisoiduissa kysymys-vastaus –järjestelmissä (esim. TREC) käytetään kuitenkin useampia taksonomioita (esim. kolmea ensimmäistä) samanaikaisesti. Pomerantz suositteleekin kaikkien viiden taksonomian hyödyntämistä tällaisten palvelujen yhteydessä. (Pomerantz 2005, 722.)

2.2.2 Kysy kirjastosta –verkkotietopalvelu

Tord Høivik (2005) on tutkinut norjalaista Kysy kirjastosta (Spør biblioteket) -verkkotietopalvelua. Se on valtakunnallinen yleisten kirjastojen verkkotietopalvelu, jota ylläpitää Oslon yleinen kirjasto (Deichmanske bibliotek). Høivikin tutkimuksessa analysoitiin 100 palveluun tullutta kysymystä. Otokseen valittiin pelkästään tietopalvelukysymyksiä; tutkimuksen ulkopuolelle rajattiin kirjastonkäyttöä koskevat neuvontakysymykset. Høivik luokitteli otoksen kysymykset kolmeen pääluokkaan:

1. Aihekysymykset (topical questions),
2. Faktakysymykset (factual questions) ja
3. Dokumenttikysymykset (document questions). (Høivik 2005, 43; 48–49.)

Aihekysymyksissä etsitään tietoa jostakin tietystä aiheesta. Ne ovat usein laajoja kysymyksiä, joihin ei voida löytää/esittää täsmällistä vastausta. Faktakysymykset ovat yksinkertaisempia kysymyksiä,

joihin on löydettävissä selkeä vastaus (esim. ’Kuka on Norjan nykyinen kuningas?’). Dokumenttikysymykset ovat faktuaalisia kysymyksiä, joissa tiedustellaan jotain tiettyä dokumenttia tai teosta (ja sen saatavuutta). Teos voi olla mm. artikkeli, kirja, runo, musiikkikappale, kuva tai elokuva. Dokumenttikysymyksiä oli Høivikin otoksessa noin kolmasosa. (Høivik 2005, 49.)

Jako fakta- ja aihekysymyksiin ei Høivikin mukaan ole aina yksiselitteinen. Ryhmien väliset rajat ovat usein likimääräisiä. Kysymys voi olla luonteeltaan faktuaalinen (esim. ’Miksi meri on suolaista?’), mutta vastaus voi vaatia laajemman aihekontekstin selvittämistä. (Høivik 2005, 49.)

Aihekysymyksiä oli Høivikin otoksessa noin puolet. Høivik jakoi aihekysymykset neljään alaluokkaan:

1) Laajat kysymykset (broad queries) (44 %),
2) Suppeat kysymykset (narrow queries) (28 %),
3) Kirjallisuuskysymykset (literary queries) (16 %) ja
4) Paikalliset kysymykset (local queries) (12 %). (Høivik 2005, 50–51.)

Laajoissa kysymyksissä käsitellään yleisesti tunnettuja aiheita (esim. kommunismi, noidat). Ne kuuluvat periaatteessa yleissivistykseen, ja niistä on helposti tietoa saatavilla. Suppeat kysymykset käsittelevät spesifimpiä aiheita (esim. Wittgensteinin väriteoriat), joista vain harvat ovat kiinnostuneita. Kirjallisuuskysymykset ovat nimensä mukaisesti kirjallisuuteen liittyviä kysymyksiä, kun taas paikalliset kysymykset käsittelevät paikallisia asioita, kuten paikallishistoriaa tai –kulttuuria. Paikalliskysymykset edellyttävät jonkin alueen paikallistietämystä. (Høivik 2005, 50–51.)

Faktakysymyksiä oli Høivikin otoksessa noin kuudesosa (18 kpl). Kolmasosa faktakysymyksistä käsitteli sanoja ja ilmauksia. Toinen kolmannes kysymyksistä liittyi luonnontieteeseen, tekniikkaan ja lääketieteeseen. Loput kysymykset liittyivät yhteiskunnallisiin asioihin ja historiaan. (Høivik 2005, 51.)

Høivikin tutkimuksessa kysymykset jaoteltiin myös Deweyn luokitusjärjestelmän pääluokkiin. Kaunokirjallisuuteen liittyviä kysymyksiä oli eniten (24 %). Toiseksi eniten tuli historiaan ja maantieteeseen liittyviä kysymyksiä (18 %). Yhteiskuntatieteisiin liittyviä kysymyksiä tuli

kolmanneksi eniten (15 %). Seuraavaksi eniten tuli tekniikkaan (13 %) ja taiteisiin ja vapaa-aikaan (12 %) liittyviä kysymyksiä. Muihin Deweyn luokitusjärjestelmän pääluokkiin tuli selvästi vähemmän kysymyksiä. (Høivik 2005, 48.)

Lisäksi Høivik tarkasteli tutkimuksessaan tiedon käyttöyhteyttä, ja sitä kautta palvelun asiakkaita. Kysy kirjastosta -palvelussa ei tiedusteltu kysyjältä kysymyksen kontekstia, mutta useimmat kysyjät kertoivat että mihin tietoa tarvitsevat. Lisäksi monissa tapauksissa kysymyksen sisältö ja sanamuoto paljastivat tiedon käyttöyhteyden. Kysymykset jaoteltiin käyttöyhteyden mukaan seuraavasti:

1. Työelämään liittyvät kysymykset,
2. kouluun ja opiskeluun liittyvät kysymykset sekä
3. epäviralliset, perhe-elämän ja vapaa-aikaan liittyvät kysymykset. (Høivik 2005, 53–54.)

Suurin osa kysymyksistä liittyi perhe-elämään ja vapaa-aikaan. Näiden kysymysten aihepiirit vaihtelivat suuresti, mutta useimmiten kysyjä tarvitsi apua jonkun dokumentin etsimisessä. Vajaa kolmannes kysymyksistä liittyi joko koulutehtäviin tai opiskeluun. Suurin osa niistä tuli koululaisilta, ja vain muutamat kysymykset tulivat korkeamman asteen opiskelijoilta. Työelämään ja työhön liittyi vain muutama kysymys, ja ne tulivat yleensä opettajilta. (Høivik 2005, 54–55.)

Høivik tarkasteli myös Kysy kirjastosta –palveluun tulleiden kysymyksien vastauksia ja niiden laatua. Kaunokirjallisuuteen liittyviin kysymyksiin vastattiin Høivikin mukaan perusteellisemmin ja laajemmin kuin luonnontieteisiin liittyviin kysymyksiin. Høivikin mukaan yleisten kirjastojen tietopalvelun laadun kannalta olisi tärkeää järjestää läheisempää yhteistyötä tiedeyhteisöjen tietopalvelujen kanssa. Høivik 2005, 56–58.)

2.2.3 Yleisten kirjastojen verkkotietopalvelututkimukset Suomessa

Yleisten kirjastojen ylläpitämiä verkkotietopalveluja (Kysy kirjastonhoitajalta, iGS) ja niihin tulleita kysymyksiä on tutkittu Suomessa toistaiseksi melko vähän. Kirjastojen omien pienimuotoisten tutkimusten ja raporttien lisäksi Suomessa on tähän mennessä tehty vain kolme laajempaa tutkimusta. Kysy kirjastonhoitajalta –palvelua ovat tutkineet Gräsbeck (2008) ja Numminen (2008). Karinen (2008) on tutkinut iGS-palvelua. Seuraavaksi tarkastellaan lähemmin näitä tutkimuksia.

Kysy kirjastonhoitajalta –verkkotietopalvelua koskevat tutkimukset

Ensimmäinen Kysy kirjastonhoitajalta –verkkotietopalvelua koskeva tutkimus tehtiin heti palvelun perustamisvuonna. Juntumaa (1998) analysoi 150 palveluun tullutta kysymystä. Kysymykset luokiteltiin seuraaviin luokkiin: 1) opastaminen, 2) aineiston paikantaminen, 3) aineiston esittely, 4) aineistoa aiheesta, bibliografisia tietoja sekä 5) asiatieto. Ihamäen (1999) tekemä jatkotutkimus käytti samaa kysymysten luokittelua. Ihamäen kyselytutkimuksessa pyrittiin tunnistamaan palvelua käyttäviä asiakasryhmiä, ja selvittämään, laajeneeko kirjaston asiakaskunta uuden palvelun myötä. Ihamäen kyselyyn vastasi 41 palvelua käyttänyttä asiakasta. Sekä Juntumaan että Ihamäen tutkimuksissa asiakkaat kysyivät eniten aineistoa aiheesta sekä asiatietoa. Nämä luokat muodostivat molemmissa tutkimuksissa noin 80 prosenttia. (Juntumaa 1998; Ihamäki 1999.)

Juntumaan (1998) ja Ihamäen (1999) tutkimukset olivat varsin suppeita. Koska ne on tehty yli kymmenen vuotta sitten, tutkimusten tulokset eivät välttämättä ole kaikilta osin ajankohtaisia. Gräsbeckin (2008) ja Nummisen (2008) tutkimukset ovat selvästi laajempia ja ajankohtaisempia, joten niitä tarkastellaan yksityiskohtaisemmin.

Tiina Gräsbeck (2008, 22) kartoitti, minkätyyppisiä kysymyksiä asiakkaat esittävät Kysy kirjastonhoitajalta palvelussa, ja millaisiin aihepiireihin kysymykset liittyvät. Tutkimuksen tavoitteena oli kysymyksiä analysoimalla ja luokittelemalla kuvata verkkotietopalvelua kirjaston palvelumuotona. Tutkimuksen aineistona oli vuoden 2007 syyskuussa palveluun tulleet kysymykset, joita oli yhteensä 426. Tutkimuksessa yksi kysymystapahtuma laskettiin yhdeksi kysymykseksi. (Gräsbeck 2008, 22–23.)

Gräsbeck jaotteli kysymykset kuuteen kategoriaan:

1. Tiettyä teosta koskevat kysymykset,
2. Kysymykset, joissa etsittiin tietoa ja aineistoa tietystä aiheesta,
3. Faktakysymykset,
4. Lukusuositukset,
5. Kirjastonkäyttöön ja kirjaston toimintatapoihin liittyvät kysymykset sekä
6. Muut kysymykset. (Gräsbeck 2008, 25–26.)

Tiettyä teosta koskevissa kysymyksissä tiedusteltiin tiettyä teosta, teoksen tekijää, tietyn tekijän teoksia ja tavallisesti myös sen saatavuutta. Pieni osa tämän kategorian kysymyksistä liittyi myös teoksen sisältöön. Myös yksittäiset artikkelit, runot, sitaatit, kuvat ja laulut kuuluivat tähän ryhmään. Lisäksi ryhmään kuuluivat kysymykset, joissa kysyttiin teoksen tekijää, tietyn tekijän teoksia, ja johonkin sarjaan liittyviä teoksia. Tiettyä teosta koskevia kysymyksiä oli tutkimuksen otoksesta kolmasosa. (Gräsbeck 2008, 25; 27–28.)

Teoskysymysten kategoriassa erottui kolme alakategoriaa. Ensimmäisessä teoksen tekijä tai teoksen nimi oli selvillä tai helposti pääteltävissä. Tällöin kysymys painottui lähinnä teoksen saatavuuteen. Näitä kysymyksiä oli reilusti yli puolet (64 %). Toisessa kategoriassa tarvittiin apua teoksen ja/tai nimen selvittämisessä. Näitä teoksen identifioimiseen liittyviä kysymyksiä oli hieman alle neljännes kaikista teoskysymyksistä. Kolmanteen alakategoriaan luokiteltiin teoksen sisältöön liittyvät kysymykset. Niissä etsittiin mm. romaanin juonitiivistelmää, käännöstä runon säkeelle tai tietoa kirjan henkilöstä. Näissä tapauksissa kysymyksen sisältö ei koskenut teoksen saatavuutta. (Gräsbeck 2008, 29–32.)

Lisäksi teoskysymysten aiheita tarkasteltiin jaottelemalla ne sen mukaan, oliko kyse kaunokirjallisuudesta, tietokirjallisuudesta, musiikkiaineistosta, video- ja DVD-elokuvista tai lehdistä ja artikkeleista. Lähes puolet (45 %) tämän ryhmän kysymyksistä liittyi kaunokirjallisuuteen. Tietokirjallisuutta tiedusteltiin toiseksi eniten (33 %), musiikkiaineistoa seuraavaksi eniten (13 %). Elokuvia tiedusteltiin toiseksi vähiten (6 %), ja vain kolme prosenttia teoskysymyksistä liittyi tiettyyn lehteen tai artikkeliin. (Gräsbeck 2008, 31–32.)

Kysymyksiä, joissa etsittiin tietoa ja aineistoa tietystä aiheesta, oli neljäsosa otoksen kysymyksistä. Aihekysymykset vaihtelivat suppeista laajoihin, mutta ne olivat yleensä laajempia kuin muiden ryhmien kysymykset. Aina kysymyksissä ei tiedusteltu aineistoa jostakin aiheesta, vaan toisinaan myös suoraa vastausta kysymykseen. (Gräsbeck 2008, 25; 27; 34.)

Aihekysymyksistä erottui kolme alakategoriaa, jotka kuvasivat kysymyksen laajuutta. Ensimmäinen alakategoria oli suppeat ja yksiselitteiset, tarkasti rajatut kysymykset (esim. ’Mistä saan rakennusoppaan autotallia varten?’). Niihin liittyvää aineistoa oli saatavilla, eikä niihin tarvinnut koota tietoa useista eri lähteistä. Toinen alakategoria oli laajemmat kysymykset, joissa kaivattiin kirjallisuutta monipuolisesti tietystä aiheesta. Kolmas alakategoria muodostui kysymyksistä, joissa aihe ei ollut vielä tarkasti rajattu, tai jota käsittelevää aineistoa ei ollut suoraan saatavilla. Tällaiset kysymykset vaativat tiedon löytämiseksi tutkimustyötä. (Gräsbeck 2008, 32–34.)

Faktakysymyksillä etsittiin yksittäistä tietoa. Niitä oli otoksessa 23 %. Faktakysymysten analyysissä kysymykset luokiteltiin YKL:n pääluokkiin. Ylivoimaisesti suurimmaksi luokaksi nousi YKL:n pääluokka 8: kaunokirjallisuus, kirjallisuustiede ja kielitiede. Faktakysymyksistä peräti 69 % kuului tähän luokkaan. Tämän luokan kysymyksissä tiedusteltiin mm. nimien alkuperää ja merkitystä, vieraskielisiä sanoja sekä kirjailijoita ja kirjallisuutta. Kaikkiin muihin luokkiin kysymyksiä tuli alle 10 %. Yksikään faktakysymys ei liittynyt filosofiaan, psykologiaan, rajatietoon, uskontoon, maantieteeseen ja kansatieteeseen. (Gräsbeck 2008, 25; 27–28; 36–37.)

Lukusuositukset –kategorian kysymyksissä pyydettiin suosittelemaan luettavaa. Kysymyksissä tiedusteltiin esim. samantyyppisiä kirjoja, joita on luettu aikaisemmin. Kysy kirjastonhoitajalta –palvelua käytettiin hyvin vähän tähän tarkoitukseen: lukusuosituksia oli nimittäin vain yksi prosentti kysymyksistä. (Gräsbeck 2008, 25; 28.)

Kirjastonkäyttöön ja kirjaston toimintatapoihin liittyvät kysymykset muodostuivat kysymyksistä, jotka liittyivät kirjaston palveluperiaatteisiin, toimintatapoihin tai lainauksenvalvontaan. Tähän ryhmään luokiteltiin myös kysymykset, jotka koskivat verkkokirjaston käyttöä sekä kysymykset, jotka koskivat kirjaston tarjoamia laitteita. Tähän kategoriaan kuului 15 % otoksen kysymyksistä. (Gräsbeck 2008, 25–26; 39.)

Viimeinen pääkategoria eli muut kysymykset, muodostui kysymyksistä joita ei voinut riittävän yksiselitteisesti luokitella edellä mainittuihin kategorioihin. Tällaisia kysymyksiä oli kolme prosenttia otoksesta. Tämän ryhmän kysymykset eivät aina liittyneet tietopalveluun lainkaan – joissain kysymyksissä tiedusteltiin mm. harjoittelu- tai työpaikkaa kirjastosta. Myös kommentit luokiteltiin tähän ryhmään. Muutama kysymys, joissa tiedusteltiin sekä tiettyä teosta että aineistoa jostakin aiheesta, luokiteltiin tähän ryhmään, koska niitä ei voitu riittävän yksiselitteisesti sijoittaa joko teos- tai aihekysymyksiin. (Gräsbeck 2008, 26–27; 42.)

Piritta Numminen (2008, 5) selvitti, minkälaisia kysymystyyppejä Kysy kirjastonhoitajalta –palvelu vastaanotti vuosina 1999 ja 2006 sekä miten kysymystyypit jakautuivat. Lisäksi Numminen vertaili kysymystyypeissä tapahtuneita määrällisiä ja laadullisia muutoksia edellä mainittujen vuosien aikana.

Tutkimuksen aineistona oli vuosina 1999 ja 2006 palveluun lähetetyt kysymykset. Vuonna 1999 palveluun lähetettiin 934 kysymystä, ja vuonna 2006 kysymyksiä tuli 2569 kappaletta. Kysymyksistä muodostettiin otos systemaattisella otannalla. Vuoden 1999 kysymyksistä otettiin joka toinen kysymys (467 kappaletta), ja vuoden 2006 kysymyksistä joka viides (513 kappaletta). (Numminen 2008, 41.)

Numminen jaotteli kysymykset neljään pääkategoriaan:

1) Neuvontakysymykset
2) Käytäntö- ja menettelytapakysymykset,
3) Ohjaavat kysymykset sekä
4) Muut kysymykset. (Numminen 2008, 31.)

Neuvontakysymykset olivat molempina vuosina suurin pääkategoria. Vuonna 1999 noin 90 prosenttia kysymyksistä oli neuvontakysymyksiä, ja vuonna 2006 niiden osuus oli 92 prosenttia. Neuvontakysymykset jaettiin kahteen alakategoriaan: 1) vaivattomiin neuvontakysymyksiin sekä 2) aiheperusteisiin tutkimuskysymyksiin. Vaivattomat neuvontakysymykset olivat faktakysymyksiä, joihin voitiin vastata nopeasti konsultoimalla ainoastaan yhtä tai kahta neuvontavälinettä. Ne edellyttivät yleensä yhtä yksinkertaista ja vaivatonta vastausta. Vaivattomien neuvontakysymysten

määrä oli lisääntynyt vuosien 1999 ja 2006 välisenä aikana 33 prosentista 45 prosenttiin. Aiheperusteisissa tutkimuskysymyksissä pyydettiin tietystä aiheesta tietyntyyppistä ja tiettyä määrää kirjoja tai lehtiartikkeleita. Niiden osuus oli hieman vähentynyt vuosien varrella: vuonna 1999 kysymyksiä oli 57 prosenttia, kun taas vuonna 2006 kysymysten osuus oli 47 prosenttia. (Numminen 2008, 21; 31; 46.)

Vaivattomat neuvontakysymykset jaoteltiin edelleen neljään alakategoriaan: 1) tunnetun nimekkeen löytäminen, 2) tunnettuun nimekkeeseen liittyvät kysymykset, 3) tietyn tai tietyntyyppisen teoksen tai aineiston löytäminen sekä 4) yksinkertaisen asian selvittäminen. Nämä neljä alakategoriaa jaoteltiin kysymysanalyysin edetessä edelleen vielä pienempiin alakategorioihin. (Numminen 2008, 47; 50.)

Myös aiheperusteiset tutkimuskysymykset jaoteltiin edelleen kolmeen alakategoriaan: 1) laajempaa selvitystyötä vaativat kysymykset, 2) tiettyyn aiheeseen liittyvät kysymykset sekä 3) tiettyyn henkilöön liittyvät kysymykset. Myös nämä kolme alakategoriaa jaoteltiin edelleen pienempiin luokkiin. (Numminen 2008, 54; 56.)

Nummisen toinen pääkategoria, eli käytäntö- ja menettelytapakysymykset, oli toiseksi suurin. Vuonna 1999 tämän ryhmän kysymyksiä oli noin kahdeksan prosenttia, ja vuonna 2006 noin seitsemän prosenttia. Tämän kategorian kysymykset luokiteltiin kahteen alakategoriaan: 1) elektronisten resurssien saatavuus ja käyttö, sekä 2) kirjaston käytöntöihin liittyvät kysymykset. Edellä mainitut alakategoriat jaoteltiin lisäksi vielä pienempiin alakategorioihin. (Numminen 2008, 31; 61–64.)

Ohjaavat kysymykset –pääkategorian kysymyksissä tiedusteltiin jonkun asian, palvelun tms. sijaintia, joko fyysisen kirjastoympäristön sisällä, tai kirjaston websivun resursseissa. Nämä kysymykset jaettiin kahteen alakategoriaan: 1) paikantaminen fyysisestä kirjastoympäristöstä, ja 2) paikantaminen kirjaston websivun resursseista. Tämän kategorian kysymyksiä oli varsin vähän: molempina vuosina noin prosentti kaikista kysymyksistä. (Numminen 2008, 31; 46.)

Muut –pääkategoriaan sisältyivät kaikki ne kysymykset jotka eivät sopineet edellä mainittuihin kategorioihin. Lisäksi tähän luokkaan sisällytettiin kaikenlaiset palautteet, kuten kiitokset ja parannusehdotukset. Tähän kategoriaan sijoittuvia kysymyksiä oli vuonna 1999 hieman yli prosentti

kaikista kysymyksistä, ja vuonna 2006 tähän kategoriaan tuli ainoastaan yksi kysymys. (Numminen 2008, 31; 46.)

Monimutkaisten neuvontapyyntöjen ennustettiin yleistyvän elektronisten informaatiopalvelujen lisääntyessä (Numminen 2008, 4). Nummisen mukaan kysymystyypeissä näyttäisi tapahtuneen täysin päinvastainen trendi kuin mitä on ennustettu. Vaivattomien neuvontakysymysten osuus oli kasvanut, ja aiheperusteisten tutkimuskysymysten vähentynyt. Nummisen mukaan aiheperusteisten tutkimuskysymysten väheneminen viittaa siihen, että yleisö käyttää hakukoneita enenevässä määrin tiettyyn aiheeseen liittyvien hakujen tekemiseen, mikä vähentää tällaisten kysymysten tiedustelemista tietopalvelusta. Nummisen mukaan kirjaston asema hankalasti selvitettävien hakutehtävien vastaajana saattaa kuitenkin säilyä, sillä laajempaa selvitystyötä vaativien kysymysten määrä oli hieman lisääntynyt. (Numminen 2008, 74.)

iGS-verkkotietopalvelua koskevat tutkimukset

iGS –verkkotietopalvelu perustettiin vuonna 2001. Tuohon aikaan kirjastossa käytiin jonkin verran keskustelua kahden samantapaisen palvelun päällekkäisyydestä ja tarpeellisuudesta. Ehkä tästä syystä ensimmäisissä iGS-palvelua koskevissa tutkimuksissa keskityttiin aluksi näiden kahden palvelun vertailuun.

Ihamäki ja Juntumaa (2002) vertasivat Kysy kirjastonhoitajalta- ja iGS-verkkotietopalvelujen kysymysten ja vastausten aihepiirejä sekä vastaamistapaa. iGS-palvelun aihepiirit jakautuivat varsin tasaisesti (kaikki aihepiirit olivat alle seitsemän prosenttia), kun taas Kysy kirjastonhoitajalta -palvelu oli selvästi ”kirjastomaisempi”. Kysy kirjastonhoitajalta –palveluun tulleista kysymyksistä noin viidennes koski kirjastonkäyttöä, ja lähes saman verran esitettiin kirjallisuuteen ja kirjailijoihin liittyviä kysymyksiä. Kysyjät ilmaisivat usein etsivänsä aineistoa jostakin aiheesta. He odottivat saavansa itse tiedon, ei niinkään tietoa aihetta käsittelevästä aineistosta. Sanojen merkitykset ja arkielämään liittyvät yllättävät tiedontarpeet olivat tyypillisiä kysymysten aiheita. (Ihamäki & Juntumaa 2002, 16–17.)

Juntumaa (2004) on lisäksi tarkastellut Helsingin kaupunginkirjastossa tarjottavia erityyppisiä verkkotietopalveluja. Verkkotietopalvelujen sisällönanalyysissään Juntumaa vertaili Kysy online

–chat neuvontaa, Kysy kirjastonhoitajalta –palvelua sekä iGS Kysy mitä vain –palvelua. Kysy online –chat neuvonta on etupäässä kirjastonkäyttöä ja HelMet-aineistohakua tukeva palvelu. Lähes puolet kysymyksistä koski HelMet-palvelua, ja myös kirjastonkäyttöön liittyviä kysymyksiä tuli runsaasti. Toiseksi eniten kirjastonkäyttöön liittyviä kysymyksiä tuli Kysy kirjastonhoitajalta –palveluun, ja vähiten iGS Kysy mitä vain palveluun. (Juntumaa 2004.)

Ville Karisen (2008) tutkimuksessa tarkasteltiin Helsingin kaupunginkirjaston ylläpitämään iGS Tietohuoltoasemaan tulleita verkkotietopalvelukysymyksiä. Tutkimuksen tavoitteena oli selvittää millaisia ovat iGS Tietohuoltoasemalle esitetyt kysymykset. Lisäksi tutkimuksessa pyrittiin rakentamaan kokonaiskuvaa iGS-palvelun käytöstä ja ominaisluonteesta. (Karinen 2008, 6.)

Tutkimusaineiston muodostivat iGS-palveluun kahden kuukauden aikana (21.11.2007 – 21.1.2008) lähetetyt kysymykset ja niihin annetut vastaukset. Tutkimus toteutettiin kysymyksiä analysoimalla ja luokittelemalla, mutta myös vastauksia analysoitiin jonkin verran. Yhteensä otokseen sisältyi 1474 kysymystä. Toisinaan asiakkaat esittivät yhteydenotoissaan useita kysymyksiä kerrallaan. Näissä tapauksissa analysoitiin vain yksi asiakkaan esittämä kysymys. Valintaperusteena käytettiin aiheenmukaista luokkaa, jonka kysymykseen vastannut Helsingin kaupunginkirjaston työntekijä oli valinnut. (Karinen 2008, 43–44.)

Karisen tutkimuksessa kysymykset luokiteltiin aluksi kolmeen pääluokkaan:

1. Varsinaiset tietopalvelukysymykset,
2. ohjaavat palvelukysymykset sekä
3. muut yhteydenotot. (Karinen 2008, 46.)

Varsinaisia tietopalvelukysymyksiä oli suurin osa (90,8 %). Muita yhteydenottoja oli 5,9 % otoksen kysymyksistä, ja kaupunginkirjastoa koskevia palvelukysymyksiä oli 3,2 %. (Karinen 2008, 46.)

Analyysin toisessa vaiheessa iGS-palveluun tulleet varsinaiset tietopalvelukysymykset luokiteltiin praktisiin ja orientoiviin kysymyksiin sekä ammattiin liittyviin kysymyksiin. Ammattiin liittyviin kysymyksiin tulkittiin myös opiskeluun ja koulunkäyntiin liittyvät kysymykset. Orientoivia

kysymyksiä oli otoksesta 63,7 %, praktisia 32,7 % ja ammattiin tai opiskeluun liittyviä vain 3,7 %. (Karinen 2008, 50.)

iGS-verkkotietopalvelussa vastaajat luokittelevat kysymykset vastaamisen yhteydessä aihealueen mukaan (61 luokkaa). Karisen tutkimuksessa näitä vastaajien tekemiä luokitteluja käytettiin iGS-palveluun tulleiden kysymysten aihealueiden kartoittamisessa. Kysymysten aiheet jakaantuivat melko tasaisesti kaikkien luokkien kesken; mikään aihepiiri ei erottunut selvästi muista. Kielet, musiikki ja kirjallisuus olivat kolme suurinta iGS:n aiheenmukaista luokkaa (yhteensä 20,3 %). (Karinen 2008, 47–48.)

Praktiset kysymykset luokiteltiin aihealueen mukaan kuuteen alaluokkaan: 1) yhteiskunta, 2) koti ja asuminen, 3) kuluttaminen, 4) terveys, 5) teknologia ja 6) muut kysymykset. Yhteiskuntaan liittyviä praktisia kysymyksiä oli eniten (22,4 %), toiseksi eniten oli kotiin ja asumiseen liittyviä kysymyksiä (20,2 %). Kolmanneksi eniten oli kuluttamiseen liittyviä kysymyksiä (16,9 %). Terveyteen liittyviä kysymyksiä oli 12,9 %, ja teknologiaan liittyviä kysymyksiä 12,7 %. Muut kysymykset olivat kysymyksiä, joita ei voinut sijoittaa mihinkään edellä mainituista praktisten kysymysten luokista. Tämän luokan kysymykset käsittelivät mm. harrastuksia, ihmissuhteita, matkailua ja muita vapaa-aikaan liittyviä aiheita. (Karinen 2008, 55–56.)

Orientoivat kysymykset luokiteltiin seuraaviin kolmeen alaluokkaan: 1) tietoa aiheesta, 2) dokumenttikysymykset ja 3) kieli ja kielitiede. Edellä mainituista luokista tietoa aiheesta –alaluokka oli suurin (58,1 % orientoivista kysymyksistä, 34 % kaikista yhteydenotoista). Toiseksi suurin alaluokka oli dokumenttikysymykset (23,5 % orientoivista kysymyksistä, 13,6 % kaikista). Näitä olivat kaikki sellaiset kysymykset, jotka liittyivät johonkin tiettyyn dokumenttiin tai teokseen. Vähiten oli kieleen ja kielitieteeseen liittyviä orientoivia kysymyksiä (18,4 % orientoivista kysymyksistä, 10,6 % kaikista kysymyksistä). Tämä luokka oli Karisen mukaan oikeastaan luokan ’tietoa aiheesta’ alaluokka, mutta se muodosti niin laajan ja sisällöltään heterogeenisen kysymysluokan, että se eroteltiin omaksi luokakseen. Orientoivien kysymysten alaluokat jaoteltiin lisäksi vielä pienempiin alaluokkiin. (Karinen 2008, 63–64; 66.)

Otoksen toinen pääluokka, palvelukysymykset, sisälsi kysymyksiä jotka liittyivät iGS-palveluun tai muihin Helsingin kaupunginkirjaston palveluihin. Ne luokiteltiin kolmentyyppisiin kysymyksiin: 1)

yleiset palvelukysymykset, 2) tekniset palvelukysymykset (verkkopalveluihin liittyvät ongelmat) sekä 3) iGS:iä tai kirjastoa koskeva palaute. (Karinen 2008, 71.)

Otoksen kolmannen pääluokan muodostivat muut yhteydenotot. Niitä ei ollut tarkoitettu varsinaisiksi kysymyksiksi. Muut yhteydenotot jaettiin seuraaviin ryhmiin: 1) asiakkaiden vastaukset iGS-kysymyksiin, mielipidekirjoitukset, 3) muut yhteydenotot. (Karinen 2008, 71–72.)

2.2.4 Hakukoneiden käyttäjätutkimuksia

Hakukoneiden käyttäjien käyttäytymistä ja hakutapoja on tutkittu mm. hakukoneiden lokitiedostojen avulla. Hakukoneiden lokitiedostoon voidaan kirjata kaikki vuorovaikutus koneen ja käyttäjän välillä. Lokitiedostojen avulla voidaan tutkia hakijan käyttäytymistä analysoimalla mm. kyselyjen ja hakuistuntojen ominaisuuksia. Lokitutkimuksen avulla voidaan saada realistista tietoa hakijoiden tavoista hakea, mutta kyselyjen taustalla olevia käyttäjän tavoitteita ja tarpeita ei lokitutkimuksissa saada selville.

Spink & Jansen (2004)

Spinkin ja Jansenin teoksessa Web search: public searching of the web (2004) tarkastellaan verkkotiedonhakua käyttäjän näkökulmasta, sekä teoreettiselta kannalta että yksityiskohtaisesti. Spink ja Jansen tarkastelevat verkkotiedonhakua omien tutkimusten kautta, joita he tekivät vuosina 1997–2004. Heidän tavoitteenaan oli selvittää, millä tavalla ”suuri yleisö” (general public) hakee tietoa verkosta, ja miten verkkotiedonhaun trendit ovat kehittyneet tänä aikana. Teoksessa on myös tiivistetty katsaus verkkotiedonhakukäyttäytymisen tutkimukseen ja tutkimusmenetelmiin vuosilta 1995–2003.

Spinkin ja Jansenin (2004) mukaan tutkimusmenetelmänä käytettiin pääasiassa kyselyjen lokianalyysia (web query transaction log analysis). Tutkimuskohteina olivat Altavistan, Exciten ja Alltheweb.comin miljoonat lokitiedostot. Tutkimuksissa tarkasteltiin mm. termien käyttöä (frekvenssiä, aiheita), kyselyitä (hakulausekkeen pituus, Boolen operaattoreiden käyttö yms.) ja hakuistuntoja (kyselyjen määrä istunnoissa, istuntojen kesto, hakutulossivujen katsominen).

Spink ja Jansen (2004) havaitsivat verkkotiedonhakijoiden kyselyjen olevan verrattain lyhyitä ja yksinkertaisia. Keskimäärin käytettiin 2.4–2.9 termiä per kysely (riippuen hakukoneesta). Edistyneempiä hakumahdollisuuksia, kuten Boolen operaattoreita, ei juurikaan hyödynnetty. Käyttö vaihteli suuresti hakukoneiden välillä, mutta keskimäärin vain 10 %:ssa kyselyistä käytettiin jotain edistyneempää hakumahdollisuutta. Eniten käytettiin fraasihakua, toiseksi eniten AND-operaattoria.

Spinkin ja Jansenin mukaan ihmiset eivät halua käyttää verkkotiedonhakuun kovinkaan paljoa aikaa ja vaivaa. Pääosa (75 %) hakuistunnoista kesti alle 15 minuuttia. Yksittäisen kyselyn osalta suurin osa verkkotiedonhakijoista tyytyi tarkastelemaan 10 ensimmäisen dokumentin tulosjoukkoa, ja tästä joukosta valittiin yleensä ainoastaan pari-kolme dokumenttia lähempään tarkasteluun. Tyypillisesti hakija tarkasteli dokumenttia noin viisi minuuttia. 15 % hakijoista tarkasteli korkeintaan 30 sekuntia.

Lucas & Topi (2002)

Yleisen käsityksen mukaan tiedonhakujärjestelmiin syötetyt kyselyt tuottavat relevantimpia tuloksia, mikäli ne sisältävät useita aiheeseen liittyviä termejä, käyttävät Boolen operaattoreita, fraasihakua ja muita edistyneempiä hakukeinoja haun täsmentämiseksi. Useissa tutkimuksissa on todettu, että yleensä internetin hakukoneisiin syötetyt kyselyt ovat lyhyitä eikä operaattoreita juurikaan käytetä. Hakutermien ja kyselyoperaattoreiden valinnan ja käytön vaikutusta hakutulosten relevanttiuteen on kuitenkin tutkittu melko vähän verkkotiedonhaun kontekstissa. Lucas ja Topi (2002) ovat selvittäneet näiden tekijöiden vaikutusta hakutuloksiin.

Tutkimukseen osallistui 87 opiskelijaa, jotka etsivät vastauksia valitsemillaan hakukoneilla kahdeksaan ennalta määriteltyyn hakuaiheeseen. Lisäksi viisi tiedonhaun ammattilaista muodosti kyselyt näistä hakuaiheista. Nämä kyselyt lähetettiin kahdeksaan hakukoneeseen, joita opiskelijat olivat käyttäneet. Hakutulosten 10 ensimmäisestä dokumentista tehtiin neliportaiset relevanssiarviot ennalta määritettyjen kriteerien mukaan. Tämän jälkeen kyselyt hakutuloksineen käytiin läpi. Jokaisen hakuaiheen parhaat kyselyt (parhaan hakutuloksen saaneet) määriteltiin ”asiantuntijakyselyiksi”, huolimatta siitä oliko kyselyn tehnyt opiskelija vai tiedonhaun ammattilainen.

Tämän jälkeen muita kyselyitä verrattiin asiantuntijakyselyyn, ja selvitettiin miten muut kyselyt erosivat asiantuntijakyselyistä hakutermien ja operaattoreiden käytön suhteen, ja millainen merkitys näillä eroavaisuuksilla oli hakutulosten suhteen. Lisäksi tutkimuksessa laskettiin ammattilaisten ja opiskelijoiden kyselyjen hakutulosten keskimääräiset relevanssiarvot, ja näitä verrattiin eri hakuaiheiden suhteen. (Lucas & Topi 2002.)

Lucasin ja Topin (2002) tutkimus vahvisti aikaisempien tutkimusten havaintoja hakutermien määrästä. Lähes 60 % kyselyistä sisälsi vain yksi tai kaksi hakusanaa. 75 % kyselyistä muodostettiin maksimissaan kolmella hakutermillä. Tärkeämpi huomio kuitenkin oli se, että nämä prosentit vaihtelivat suuresti eri hakuaiheiden välillä, sillä hakuaiheella on suuri merkitys käytettyjen hakutermien määrään. Tämän lisäksi myös hakuaihe vaikutti merkittävästi käytettyjen operaattoreiden määrään. Monimutkaisimmissa hakuaiheissa operaattoreita käytettiin yli 50 % kyselyistä.

Lucasin ja Topin (2002) tutkimuksessa havaittiin myös, että tiedonhaun ammattilaiset käyttivät enemmän operaattoreita kuin opiskelijat. Tämä tukee aikaisempien tutkimusten havaintoja. Ammattilaiset käyttivät opiskelijoita enemmän hakutermejä, mutta ero ei ollut niin merkittävä kuin operaattorien käytön suhteen. Myös Hsieh-Yee (1993) ja Hölscher & Strube (2000) ovat havainneet, että noviisien ja asiantuntijoiden käyttämien hakutermien määrä ei välttämättä vaihtele kovinkaan paljon. Lucas ja Topi havaitsivat kuitenkin, että tarkasteltaessa hakutermien määriä hakuaiheittain, ammattilaisten ja opiskelijoiden väliltä löytyi selviä eroja. Lucasin ja Topin mukaan tällaiset keskimääräiset arviot hakutermien määrästä tai operaattoreiden käytöstä voivat olla harhaanjohtavia, mikäli kyselyjä ei tarkastella hakutehtävän kontekstissa.

Lucas ja Topi (2002) havaitsivat, että hakutermien valinnalla ja käytöllä oli paljon suurempi merkitys kyselyn onnistumiseen kuin operaattoreiden valinnalla ja käytöllä. Samansuuntaisiin tuloksiin päätyi myös Jansen (2000). Mikäli hakija ei ole tyytyväinen hakutulokseen, aivan pienetkin muutokset voivat johtaa merkittävästi erilaisiin – ja mahdollisesti parempiin – tuloksiin.

Hölscher & Strube (2000)

Hölscher ja Strube (2000) ovat tutkineet tietämyksen ja verkkotiedonhaun yhteyttä. Heidän tutkimuksessaan selvitettiin millaisella tietämyksellä on merkitystä verkkotiedonhaun kannalta, ja millaisia tietämysrakenteita ja strategioita verkkotiedonhakuun liittyy. Samalla he tutkivat tiedonhaun ”noviisien” ja ”asiantuntijoiden” hakukäyttäytymistä. Hölscher ja Strube lähestyivät ongelmaa kaksiosaisen tutkimuksen kautta. Ensimmäisessä tutkimuksessa haastateltiin 12 tiedonhaun ammattilaista, ja kartoitettiin heidän hakustrategioitaan ja hakukäyttäytymistään. Tämän jälkeen haastateltavat suorittivat joukon hakutehtäviä valitsemallaan hakukoneella – eivät kuitenkaan itse, vaan neuvomalla suullisesti avustajiaan, ja kertomalla samalla ääneen omia tiedonhakuprosessiin liittyviä ajatuksiaan. Ensimmäisen tutkimuksen avulla Hölscher ja Strube rakensivat verkkotiedonhaun mallin, jota he testasivat toisessa tutkimuksessa.

Hölscherin ja Struben (2000) toisessa tutkimuksessa testattiin tiedonhakuun liittyvän kokemuksen ja aihekohtaisen tietämyksen merkitystä tiedonhaun kannalta. 24 osallistujaa jaettiin neljään kuuden hengen ryhmään heidän tiedonhakukokemuksensa (vähän/paljon) ja aihekohtaisen tietämyksen suhteen (vähän/paljon). Kukin osallistuja suoritti sarjan vaativia talouteen liittyviä hakutehtäviä. Proxy-lokitiedostoon tallennettiin mm. kaikki hakijoiden suorittamat kyselyt sekä käydyt URL-osoitteet.

Tutkimuksen ensimmäisessä vaiheessa 12 asiantuntijahakijaa käytti keskimäärin 3,64 hakusanaa kyselyä kohti. Esimerkiksi saksalaisen Fireball-hakukoneen käyttäjien kyselyissä hyödynnettiin vain 1,66 hakusanaa per kysely. Tutkimuksen toisessa osassa, missä verrattiin ”noviisien” ja ”asiantuntijoiden” välisiä eroja, Hölscher ja Strube havaitsivat että tiedonhaun asiantuntijat käyttivät noviiseihin verrattuna paljon useammin edistyneempiä hakuvaihtoehtoja, kuten Boolen operaattoreita, fraasihakua jne. Hakusanojen määrän suhteen ero ei ollutkaan enää kovinkaan suuri: 12 asiantuntijahakijan kyselyt sisälsivät keskimäärin 2,61 sanaa, ja noviiseilla 2,32 sanaa. Neljä asiantuntijaa, joilla oli hyvä aihetietämys, käyttivät ainoastaan 1,97 sanaa kyselyä kohden, kun taas asiantuntijat, joilla oli vain vähän aihetietämystä, käyttivät kyselyissä keskimäärin 2,96 sanaa. Lisäksi neljä noviisitiedonhakijaa, joilla oli hyvä aihetietämys, kompensoivat kyselyn muodostamistaitojensa puutteita suuremmalla verbaalisella luovuudella. Eniten ongelmia oli noviisitiedonhakijoilla, joilla ei ollut hyvää aihetietämystä. Parhaiten hakutehtävistä selvisi tiedonhaun asiantuntijoiden ryhmä, jolla oli lisäksi hyvä aihetietämys. Tästä voidaan päätellä, että

onnistuneet verkkotiedonhaut perustuvat sekä tiedonhakukokemuksen määrään että aihetietämykseen. Lisäksi nämä löydökset tukevat sitä käsitystä, että kyselyjen termien valinnalla on huomattavasti enemmän merkitystä kuin käytettyjen termien lukumäärällä.

Jansen (2000)

Jansen (2000) selvitti kyselyn rakenteen vaikutusta verkkotiedonhaun tuloksiin. Hänen tutkimuksessaan valittiin Excite-hakupalvelun kyselylokitiedostoista 15 yksinkertaista kyselyä (ilman Boolen operaattoreita, fraasihakua jne.). Nämä kyselyt lähetettiin viiteen suureen hakukoneeseen (Alta Vista, Excite, FAST Search, Infoseek, Northern Light). Kyselyjen hakutuloksista muodostettiin tutkimuksen perusjoukko. Jansenin tutkimuksessa hakutuloksista poimittiin ainoastaan 10 ensimmäistä dokumenttia.

Tutkimuksen seuraavassa osassa näitä yksinkertaisia kyselyitä muokattiin monimutkaisemmiksi hakukoneiden pääsivulta löytyvillä hakuvaihtoehdoilla. Kysymyksiin mm. lisättiin Boolen operaattoreita ja +/- operaattoreita sekä tehtiin fraasihakuja. Kyselyitä muodostettiin yhteensä 150, ja ne lähetettiin uudestaan edellä mainittuihin hakukoneisiin. Saatuja hakutuloksia verrattiin yksinkertaisista kyselyistä saatuun perusjoukkoon. Vertailussa tarkasteltiin ainoastaan dokumenttien päällekkäisyyksiä hakutuloksissa, mitään relevanssiarviota dokumenteista ei tehty. Tutkimuksessa käytiin läpi yhteensä 2 768 hakutulossivua.

Jansenin (2000) tutkimus osoitti melko selvästi, ettei kyselyn kompleksisuuden lisäämisellä ollut suurta vaikutusta hakutuloksiin. Verrattaessa yksinkertaisten ja kompleksisempien kyselyjen hakutulosten eroja havaittiin että keskimäärin yli 70 % hakutuloksien dokumenteista vastasi toisiaan. Tutkimuksessa havaittiin myös, ettei millään tietyn operaattorin käytöllä ollut merkittävää vaikutusta hakutuloksiin. Jansenin mukaan tyypilliset verkkotiedonhakijat, jotka tekevät lyhyitä ja yksinkertaisia kyselyjä hakukoneisiin, käyttäytyvätkin melko lailla järkevästi. Jos kyselyn kompleksisuuden lisäämisellä ei ole kovinkaan suurta merkitystä hakutuloksen kannalta, voidaan kysyä miksi kannattaisi vaivautua opettelemaan edistyneempiä hakutaktiikoita? Jansenin mukaan tällaiset tulokset saattavat johtua verkkotiedonhakujärjestelmien ranking-algoritmeista, jotka tukevat tyypillisen keskivertokäyttäjän hakutapoja.

3. TUTKIMUSASETELMA

3.1 Tutkimuksen tarkoitus ja tutkimusongelmat

Tutkimuksen kohteena on Helsingin Kaupunginkirjaston iGS-verkkotietopalvelu ja siihen lähetetyt kysymykset. Tutkimuksen päätavoitteena on selvittää, että kuinka hyvin Google-hakukoneen keskivertokäyttäjien on mahdollista löytää vastauksia iGS-palveluun lähetettyihin kysymyksiin. Tutkimuksen toisena tavoitteena on selvittää että minkä tyyppisiä kysymyksiä palveluun lähetetään. Tarkoituksena on kysymyksiä analysoimalla ja luokittelemalla lisätä ymmärrystämme verkkotietopalveluista ja niiden kysymystyypeistä. Kolmantena tavoitteena on selvittää että iGS-palveluun lähetettyjen kysymyksien tyypillä yhteyttä Google-hakukoneen avulla saatuihin hakutuloksiin. Käytännössä selvitetään, onko tietyntyyppisiin kysymyksiin helpompaa tai vaikeampaa löytää vastauksia Googlella.

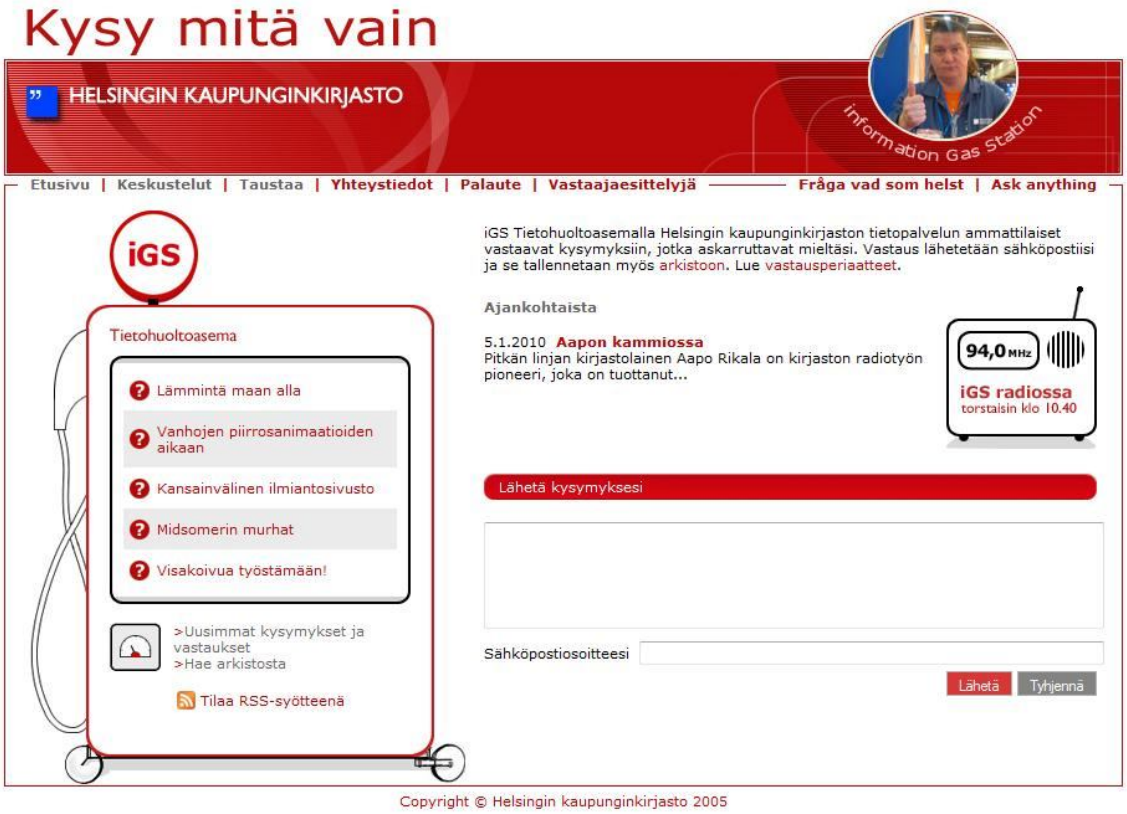
Tutkimusongelmat ovat:

1. Minkätyyppisiä kysymyksiä asiakkaat lähettävät iGS-verkkotietopalveluun?
2. Missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin?
3. Onko iGS:n kysymysten tyypillä yhteyttä siihen että missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin?

3.2 Tutkimuksen kohde: iGS Tietohuoltoasema

iGS Tietohuoltoasema (information Gas Station). on Helsingin kaupunginkirjaston ylläpitämä, maksuton verkkotietopalvelu. iGS–projekti käynnistyi kesällä 2000, kun Helsingin kaupunginkirjasto vastaanotti Bill & Melinda Gatesin säätiön myöntämän Access to Learning –palkinnon. Osa palkintorahoista käytettiin innovatiivisen Tietohuoltoasema-kokeiluprojektin käynnistämiseen. Kokeiluprojekti suunniteltiin alun perin kaksivuotiseksi (2001–2002), mutta iGS:n toiminta jatkuu edelleen. IGS verkkotietopalvelu on muuttunut projektista kirjaston vakiintuneeksi palveluksi. (iGS Tietohuoltoaseman projektiraportti 2001–2002.)

iGS:n verkkopalvelussa asiakkaat voivat lähettää kysymyksiä hyödyntämällä kuviossa 1 esitettyä käyttöliittymää. Palvelu on maksuton, eikä se vaadi rekisteröitymistä. Vastaus lähetetään asiakkaan sähköpostiin, mikäli sellainen on ilmoitettu. Kysymyksen voi jättää myös ilman sähköpostiosoitetta. Tällöin vastauksen voi lukea iGS:n verkkosivuilta.



Kuvio 1. iGS-verkkotietopalvelun käyttöliittymä (http://igs.kirjastot.fi)

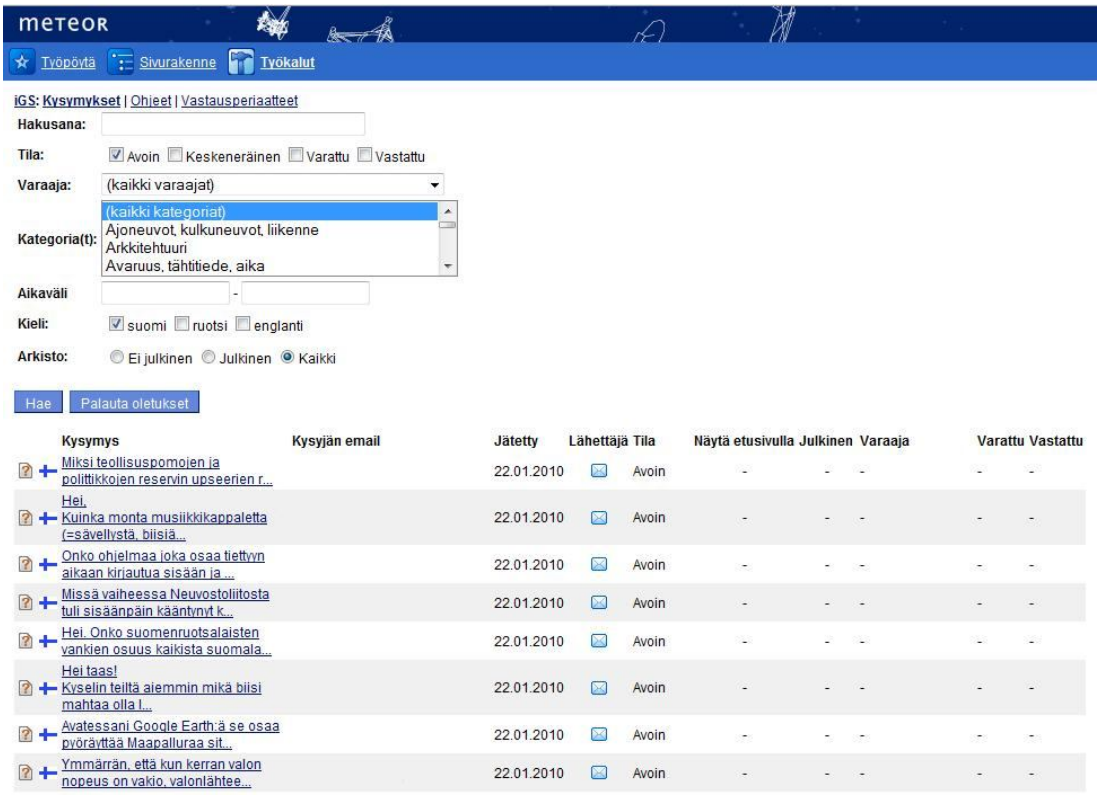
Kysymykset vastauksineen tallennetaan julkiseen arkistoon, joten ne ovat kaikkien luettavissa. Arkisto sisältää vastauksia vuodesta 2001 lähtien. Osa kysymyksistä voidaan jättää julkaisematta (mm. asiattomat viestit tai erityisen henkilökohtaiset kysymykset). Vastaukset asiasanoitetaan ja luokitellaan sisällön osalta kategorioihin (62 kpl). iGS:n verkkosivuilla näkyy 3000 viimeisintä kysymystä aikajärjestyksessä. Lisäksi kysymyksiä voi hakea iGS:n arkistosta hakusanoilla, tai selailla kategorioittain. iGS:n julkiseen arkistoon oli vuoden 2007 loppuun mennessä kertynyt 32328 kysymystä. Ei-julkisia kysymyksiä oli 2388 kappaletta. (iGS Tietohuoltoaseman projektiraportti 2001–2002; Juntumaa 2003; Granlund & Lönn 2007.)

iGS tietopalvelussa kirjastomaisuus on jätetty tarkoituksella taka-alalle. Se on tarkoitettu nopeiden, käytännöllisten ja usein ”ei-kirjastomaisten” tiedontarpeiden kanavaksi. Kysymyksiä esitetään laidasta laitaan, eikä mikään yksittäinen aihe erotu selvästi joukosta. Tyypillisiä aiheita ovat esimerkiksi sanojen merkitykset, arkielämään liittyvät yllättävät tiedontarpeet ja henkilökohtaiset pulmat. (Tietotekniikka- ja verkkopalvelustrategia 2003–2006; Ihamäki & Juntumaa 2002.)

iGS-palvelussa pyritään antamaan asiakkaalle vastaus kahden viikon sisällä. Vastauksissa käytetään myös lähdeviitteitä, mutta viitteiden lisäksi pyritään aina antamaan myös suora vastaus kysymykseen. Tiedon ja tiedonlähteiden tarjoamisen lisäksi iGS:n tavoitteena on opastaa asiakkaita tiedonhankinnassa ja lähdekritiikissä. Vastauksia etsitään mm. internetistä, kirjallisuudesta, tietokannoista sekä asiantuntijoilta. iGS-palvelussa ei tehdä mitään syvällistä tutkimusta, vaan vastaukset pyritään pitämään yksinkertaisina ja selkeinä. Vastaustapa ja vastauksen laajuus riippuu paljon vastaajasta, mutta tämä on tarkoituskin: iGS-palvelussa pidetään tärkeänä erilaisia, persoonallisia vastaustapoja. Palvelun tehtäväksi on nähty paitsi tiedon myös ilon tuominen kysyjien ja arkiston käyttäjien elämään. Tästä syystä vastaukset saattavat olla esimerkiksi pohdiskelevia, kolumnityylisiä tai humoristisia, kysymyksen luonteesta ja vastaajasta riippuen. (Granlund & Lönn, 2007.)

iGS:n tiimin muodostavat kysymyksiin muun työnsä ohessa vastaavat Helsingin kaupunginkirjaston työntekijät. iGS:n vastaajaverkossa oli vuonna 2007 mukana 120 kirjaston työntekijää. Verkostossa oli jäseniä 28 toimipisteestä. Aktiivisin vastaajajoukko (vastanneet yli 30 kertaa vuoden aikana) koostui 48 vastaajasta. Vuoden 2007 aikana vastattiin 7560 kysymykseen, ja verkkokäyntejä sivuilla oli lähes 500 000. (iGS Tietohuoltoaseman toimintakertomus 2007.)

Kysymyksiin vastataan alla näkyvän www-pohjaisen käyttöliittymän kautta (Kuvio 2). iGS:n verkkosivujen kautta lähetetyt kysymykset ilmaantuvat reaaliaikaisesti käyttöliittymän sivuilla näkyvälle listalle. Vastaajat voivat varata itseään kiinnostavan kysymyksen, ja jatkaa kysymykseen vastaamista silloin kun itselle parhaiten sopii.



Kuvio 2. iGS-verkkotietopalvelun virkailijaliittymä (http://www.lib.hel.fi/administration/logon.aspx)

3.3 Tutkimuksen väline: Google

Google on maailman suosituin ja tunnetuin hakukone. Larry Page ja Sergey Brin perustivat Google-yrityksen syyskuussa vuonna 1998. Googlen käyttämän hakualgoritmin yksityiskohdat ovat tarkasti varjeltuja salaisuuksia. Googlen hakumoottori perustuu ainakin kahteen ominaisuuteen: täydelliseen täsmäytykseen ja PageRank-järjestelmään. Täydellinen täsmäytys perustuu merkkijonojen täydelliseen täsmäytykseen ja Boolen operaattorien käyttöön. Täydellisessä täsmäytyksessä hakutulokseen etsitään vain sellaiset dokumentit, joista löytyy hakulausekkeessa ilmaistut sanat täsmälleen siinä muodossaan. Nykyisin Google käyttää myös jossain määrin automaattista sanojen stemmausta (sanojen pääteaineisten karsiminen) hyväkseen, joten se ymmärtää hakutermin eri taivutusmuotoja. Google ei tue katkaisumerkkiä, joten hakusanan muodolla on suuri merkitys haun onnistumisen kannalta. Lisäksi Googlen perushaku palauttaa hakiessa vain ne sivut, jotka sisältävät kaikki hakutermit. Boolen ”AND” –operaattori on kaikissa yli yhden hakusanan hauissa aina mukana (ellei hakija itse yhdistele hakusanoja muilla Boolen operaattoreilla). Google-haut eivät

erottele isoja ja pieniä kirjaimia; kaikki kirjaimet käsitellään hauissa pieninä kirjaimina. (Googlen verkkosivut.)

Googlen eräs menestyksen salaisuus on Larry Pagen ja Sergey Brinin kehittämä sivujen rankkauksen PageRank-järjestelmä. PageRank käyttää netin linkkirakennetta hyväkseen tietyn sivun arvioimisessa. Sitä voisi kutsua myös jonkinlaiseksi ”nettidemokratiaksi” (tosin hieman epätasa-arvoiseksi), missä sivuja rankataan sen mukaan kuinka paljon ”ääniä” ne saavat muilta verkkosivuilta. Google laskee sivulta A sivulle B vievän linkin yhdeksi ääneksi sivun B hyväksi. Sellaiset sivut, joihin on paljon linkkejä, nousevat listalla parempaan asemaan. Lisäksi Google analysoi myös sivun, joka antaa äänen. ”Tärkeiden” sivujen antamilla äänillä on enemmän painoarvoa ja ne näin auttavat tekemään muista sivuista ”tärkeitä”. Google käyttää internetin ”kollektiivista älykkyyttä” arvioidessaan sivujen ”tärkeyttä” – ihmiset eivät osallistu sivujen arviointiin. Tärkeät korkealaatuiset verkkosivut saavat korkeamman PageRank –arvon, jonka Google muistaa jokaisella hakukerralla. (Googlen verkkosivut.)

Tässä tutkimuksessa käytetään Googlen etusivulla näkyvää perushakua. Googlen etusivulla on myös linkki Tarkennettu haku, missä voi tehdä erilaisia hakuun liittyviä rajauksia. Tutkimuksessa oletetaan, että hakukoneiden peruskäyttäjä ei käyttäisi Googlen Tarkennettua hakua.

3.4 Tutkimuksen metodista

Tutkimuksen pääongelmaa, ”missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin?”, voisi periaatteessa lähestyä esimerkiksi teettämällä hakutehtäviä testiryhmällä. Tällainen lähestymistapa olisi tämän tutkimuksen kannalta ongelmallinen. Valitun testiryhmän tulisi olla lukumäärältään varsin suuri, jotta se edustaisi tilastollisesti hakukoneiden keskivertokäyttäjiä. Tällaisen ryhmän kokoaminen ja testiolosuhteiden järjestäminen olisi liian vaativa prosessi tämänkokoisen tutkimuksen kannalta.

Toinen lähestymistapa, tekemällä kyselyitä itse omia hakutaitoja ja mielikuvitusta käyttäen, on luonnollisesti aivan liian subjektiivinen tapa tutkia aihetta. Kolmas vaihtoehto, jota tässä tutkimuksessa sovelletaan, on hakukoneiden keskivertokäyttäjän ”simulointi” iGS-kysymysten avulla.

Hakukoneiden keskivertokäyttäjän ”simulointi” voi vaikuttaa varsin erikoiselta lähestymistavalta. Vaikka saisimme kerättyä valtavia määriä tietoa hakukoneiden keskivertokäyttäjistä, haun toteuttaminen on kuitenkin aina hyvin subjektiivinen prosessi. Ihmiset ilmaisevat tiedontarpeitaan hyvin erilaisilla käsitteillä, ja käsitteitä taas ilmaistaan lukuisin eri tavoin merkkijonotasolla. Kun hakutermejä voidaan vielä yhdistellä lukuisin eri tavoin (esim. Boolen operaattoreilla), on selvää että yksinkertainenkin hakupyyntö voidaan esittää kyselynä monin eri tavoin. Ei ole olemassa ehdottomia kriteerejä, joiden mukaan voisi sanoa että keskivertokäyttäjä muotoilisi jonkin hakupyyntöön liittyvän kyselyn tietyllä tavalla.

Hakukoneiden keskivertokäyttäjän ”simulointi” on silti sopiva lähestymistapa tämän tutkimuksen kannalta. Hakukoneiden keskivertokäyttäjistä saatua tutkimustietoa voidaan nimittäin soveltaa eräänlaisten tutkimuksen ennakkoehtojen ja rajausten luomisessa. Täydellistä keskivertokäyttäjän ”simulaatiota” ei pystytä tekemään, mutta tutkimustulosten avulla on mahdollista muodostaa sellaisia hakulausekkeiden muotoilemisen sääntöjä, jotka pyrkivät vähentämään hakijan subjektiivisten ominaisuuksien, kuten tiedonhakutaitojen ja mielikuvituksen, vaikutusta hakutuloksiin. Näistä ”simulaatioon” liittyvistä säännöistä kerrotaan tarkemmin luvussa 3.7.

Tutkimusmetodi käydään yksityiskohtaisesti läpi seuraavissa luvuissa. Metodi on tiivistettynä seuraavanlainen: Aluksi otetaan otos iGS-verkkotietopalvelun kysymyksistä. Otoksen hakupyynnöt eli iGS-kysymykset analysoidaan ja luokitellaan sopiviin kategorioihin, ja tutkimuksen kannalta turhat hakupyynnöt karsitaan pois. Aineiston käsittelystä on kerrottu tarkemmin luvussa 3.5. Tämän jälkeen otoksen hakupyynnöille tehdään käsitteellinen analyysi, jonka jälkeen käsitteistä muodostetaan hakukoneiden keskivertokäyttäjien ”simulaation” avulla hakulausekkeita Googlea varten. Käsitteellisestä analyysistä ja ”simulaation” säännöistä kerrotaan tarkemmin luvuissa 3.6 ja 3.7. Google-hauissa löytyneille dokumenteille tehdään relevanssiarvio (ks. luku 3.8), jonka jälkeen hakutulokset (löytyi / ei löytynyt vastausta), hakupyynnöt, hakupyynnöistä löytyneet käsitteet, kyselyt ja relevanttien dokumenttien URL-osoitteet kirjataan muistiin. Lopuksi aineisto käsitellään ja analysoidaan käyttäen apuna tilastollista ohjelmistoa (SPSS). Aineiston analyysistä on kerrottu tarkemmin luvussa 3.9.

3.5 Aineiston käsittely

iGS-verkkotietopalvelun arkistosta (http://igs.kirjastot.fi/iGS/kysymykset/haku.aspx) löytyy 3000 viimeisintä tietopalvelukysymystä. Kysymykset ovat aikajärjestyksessä (uusimmat etusivulla), joten arkistosta pystyy helposti poimimaan jonkun tietyn aikavälin kysymykset otokseksi.

Tutkimusta varten poimittiin kolmen viikon iGS-kysymykset vastauksineen vuodelta 2007. Jotta otoksen ajankohdan mahdollinen vaikutus tutkimuksen tuloksiin voitaisiin minimoida, otokseen valittiin viikot 2, 12 ja 22. Tällä tavalla otoksessa on edustettuna kaikki viikonpäivät, kuukauden eri ajankohtia (alkupuoli, puoliväli ja loppu) sekä vuodenaikoja (talvi, kevät ja kesä):

- Viikko 2 (8.1.–14.1.2007): kuukauden alkupuoli talvi
- Viikko 12 (19.3.–25.3.2007): kuukauden puoliväli kevät
- Viikko 22 (28.5.–3.6.2007): kuukauden loppu kesä

Näiltä kolmelta viikolta kertyi yhteensä 382 iGS-arkiston kysymystä. Poimitut kysymykset kopioitiin tekstinkäsittelyohjelmaan. Kysymykset numeroitiin aikajärjestyksessä (1–382).

Kysymyksiin tarkemmin tutustuessa kävi ilmi, että yksittäinen kysymys saattoi sisältää useampia kysymyksiä, joko samaan aiheeseen liittyen, tai kokonaan eri aihetta koskevia. Nämä kysymykset jaettiin alakysymyksiin, jotka muotoiltiin omin sanoin selkeämpään muotoon. Alakysymyksiä kertyi yhteensä 530 kappaletta, joten yhtä iGS-kysymystä kohden tuli keskimäärin 1,4 alakysymystä.

Kysymysten käsittelyn helpottamiseksi alakysymyksille annettiin kysymyksen numeron lisäksi aakkosista muodostuva loppuosa. Lisäksi alkuperäiset kysymykset ja alakysymykset eroteltiin eri väreillä. Alla olevassa esimerkissä alkuperäinen kysymys on tumman harmaalla pohjalla, ja alakysymykset vaalean harmaalla pohjalla:

(09.01.2007) **Hei! Tutkin vanhoja kauppakirjoja ja niissä oli sana tollikka, mitä se tarkoittaa, se oli maininta tilalla olevista rakennuksista ja niiden kuulumisesta kauppaan. ...niemitollikko ja hakotollikko. Kauppakirja oli vuodelta 1936. Asun tällä ko. paikalla Kolpeneenharjulla Rovaniemellä. Onko sana ehkä saamenkielinen?**
31A: Mitä tarkoittaa vanhoissa kauppakirjoissa esiintyvä sana tollikka? (F) 31B: Onko sana tollikka saamenkielinen? (F)

Alakysymyksiin jaottelu ja kysymysten uudelleen muotoilu oli välttämätöntä monesta syystä. Ensinnäkin hakujen onnistumista pystyttiin tarkastelemaan täsmällisemmin; usein kävi nimittäin niin, että vastausta ei löytynyt Googlella kuin johonkin alakysymykseen. Toiseksi hakulausekkeita oli helpompi muodostaa kun oli selkeitä yksittäisiä kysymyksiä joihin täytyi etsiä vastausta. Kysymyksiin jaottelu oli myös välttämätöntä kysymyksen ymmärtämisen kannalta. Asiakkaiden lähettämät kysymykset olivat usein melko pitkästi ja vaikeasti ilmaistuja, ja toisinaan oli vaikeaa ymmärtää mitä asiakas todella haluaa tietää. Läheskään aina ei voitu kirjoittaa asiakkaan esittämää kysymystä sellaisenaan, vaan piti aina ensin ottaa huomioon konteksti, eli muut kysymyksessä ilmenevät asiat. Tästä syystä alakysymyksiin jaottelu ja kysymyksen uudelleen muotoilu ei ollut aina ongelmatonta.

Alakysymyksiä muotoillessa piti myös päättää, kuinka monta alakysymystä alkuperäisestä iGS-kysymyksestä voi muodostaa. Tämäkään ei aina ollut yksinkertaista. Kysymyksiä voi muotoilla monin eri tavoin, ja asioita voi yhdistää samaan kysymykseen, tai sitten ne voi pitää erillisinä kysymyksinään. Toisinaan erottelu oli välttämätöntä, joskus oli parempi yhdistellä kysymyksiä. Esimerkiksi kysymys numero 200:

(22.03.2007) **Eri maissa tervehdyssuudelmat poikkeavat toisistaan. Missä maissa käytetään tervehdyssuudelmia, ja "miten", kenelle, koska...? Kuinka läheiset ihmiset tervehtivät toisiaan suutelemalla?** **Ellei muuta niin haluaisin tietää mistä löytäisin tietoa näistä asioista.**
200: Haluaisin tietoa tervehdyssuudelmista (missä maissa käytetään, miten se tapahtuu, kenelle se tapahtuu jne.) ja löytyisikö mahdollisia lähteitä? (A)

Vaikka kysyjä esittää useita erillisiä kysymyksiä, kaikki kysymykset koskevat tervehdyssuudelmia. Kysyjä haluaa mahdollisimman paljon tietoa tervehdyssuudelmista, ja mahdollisia lähteitä aiheesta. Tästä syystä erilliset kysymykset yhdistettiin yhdeksi kysymykseksi.

Alakysymyksiin jaon jälkeen kysymykset luokiteltiin seuraaviin kategorioihin:

1. **Tiettyä teosta koskevat kysymykset**

   Tähän ryhmään kuuluvat kysymykset, joissa tiedustellaan tiettyä teosta ja mahdollisesti myös sen saatavuutta. Yksittäiset artikkelit, runot, sitaatit, kuvat ja laulut kuuluvat myös tähän ryhmään. Tämän ryhmän kysymyksissä saatetaan myös kysyä teoksen tekijää, tietyn tekijän teoksia ja johonkin sarjaan liittyviä teoksia. Kysymykset saattavat liittyä myös

teoksen identifioimiseen (esim. juonikuvauksen perusteella) tai sen sisältöön (esim. tietoa kirjan henkilöstä).

2. **Tiettyä aihetta koskevat kysymykset**
   Tähän ryhmään kuuluvat kysymykset, joissa etsitään tietoa ja/tai aineistoa tietystä aiheesta. Aihekysymykset ovat yleensä laajempia kuin ryhmien 1 ja 3 kysymykset. Kysymyksissä saatetaan tiedustella kirjallisuutta, nettisivuja tai muuta aineistoa jostakin aiheesta. Lisäksi ryhmään kuuluvat kysymykset, joissa halutaan laajempi vastaus tai selvitys tiettyyn aiheeseen liittyen.

3. **Faktakysymykset**
   Tähän ryhmään kuuluvat kysymykset, joissa kysytään yksittäistä tietoa. Faktakysymyksissä voidaan kysyä esimerkiksi nimiä, vuosilukuja, sijainteja, määriä, eli tarkasti määriteltyjä faktoja. Tähän ryhmään kuuluvat myös määritelmät sekä sanojen/sanontojen alkuperää koskevat kysymykset, mikäli vastaus on riittävän yksiselitteinen. Moniselitteisen vastauksen tapauksessa kysymys luokitellaan tiettyä aihetta koskevaksi kysymykseksi.

4. **Muut kysymykset**
   Tähän ryhmään kuuluvat kysymykset, joita ei voi luokitella riittävän yksiselitteisesti kategorioihin 1–3.

Kategoriat merkittiin alakysymysten perään seuraavasti: (T) – Tiettyä teosta koskevat kysymykset, (A) – Tiettyä aihetta koskevat kysymykset, (F) – Faktakysymykset sekä (M) – Muut kysymykset.

Kysymyksiä analysoitaessa muut kysymykset –kategorian kysymykset osoittautuivat lopulta sellaisiksi, ettei niitä ollut mahdollista käyttää tutkimuksen analyysivaiheessa. Tämän ryhmän ”kysymykset” olivat mm. kommentteja aikaisempiin kysymyksiin, huumorikysymyksiä joihin ei odotettu järkevää vastausta (esim. ”Tuli tossa mieleen…Olikohan se Nooan arkin koko A4 vai A5?”) sekä HelMet–aineistohakuun liittyviä kysymyksiä (joihin ei voi Googlella etsiä vastausta). Lisäksi suomenkielisten kysymysten joukkoon oli eksynyt muutama vieraskielinen kysymys, vaikka palvelussa on mahdollista lähettää kysymys ruotsin- ja englanninkielisellä käyttöliittymällä. Tämän kategorian kysymyksiä tuli yhteensä 21 kappaletta, ja ne karsittiin pois tutkimuksen analyysivaiheessa. Tämän jälkeen analyysiin sopivia alakysymyksiä oli yhteensä 509 kappaletta.

3.6 Hakuaiheen käsiteanalyysi

Kyselyn muotoilu etenee useimmiten – tietoisesti tai tiedostamatta – tasoittain käsitetasolta ilmaisutasolle ja siitä merkkijonotasolle. Ensiaskel kyselyn laadinnassa on siis tunnistaa hakuaiheen keskeiset käsitteet ja käsitteiden väliset suhteet, eli selvittää hakutehtävän käsitteellinen rakenne. Analyysin tuloksena on käsitteellinen hakusuunnitelma, jonka pohjalta voidaan muodostaa myöhemmin toteutettava kysely. Käsitteellinen hakusuunnitelma kiinnittää huomion siihen, *mitä* tietoa haetaan ja kysely siihen, *miten* sitä haetaan. Haun suunnittelun tavoitteena on kysely, jossa esiintyy hakuavaimia. Tähän tavoitteeseen edetään hakuaihetta kuvaavien käsitteiden kautta. (Järvelin 1995, 142.)

Käsitteellisen hakusuunnitelman sisältämät käsitteet jäsennetään rinnakkaisiin ja rajaaviin suhteisiin sen mukaan, edustavatko käsitteet samoja vai eri aspekteja hakuaiheessa. Aspekti on hakuaiheeseen liittyvä näkökulma, fasetti, joka on suhteena väljempi kuin käsitteiden hierarkkinen suhde – ne voivat olla keskenään esimerkiksi assosiaatiosuhteessa. Suhteet ovat rinnakkaiset, jos käsitteet edustavat hakusuunnitelmassa samaa aspektia ja muulloin rajaavat. (Järvelin 1995, 142–144.)

Kun hakuaiheen aspektit on valittu, tulee tunnistaa aspekteja edustavat käsitteet. Yhtä hakuaiheen aspektia kohti voi löytyä useita hierarkkisesti eritasoisia käsitteitä. Osa käsitteistä saadaan suoraan hakutehtävästä, mutta osa joudutaan yleensä päättelemään. (Järvelin 1995, 147.) Tässä tutkimuksessa kyselyjen muodostamisessa käytetään pelkästään suoraan hakutehtävästä saatavia käsitteitä. Tämä rajaus on välttämätön tutkimusasetelman kannalta.

Hakukäsitteiden valinta ei ole hakupyynnön automaattista pilkkomista osiin, vaan valintaa mukaan otettavista ja poisjätettävistä käsitteistä. Hakuaihe saattaa sisältää esimerkiksi käsitteitä, jotka ovat haun tuloksellisuuden kannalta turhia tai jopa haitallisia. Käsitteiden tunnistamisessa tulee pohtia, että mitkä ovat hakuaiheen kannalta tärkeimpiä käsitteitä, ns. pääkäsitteitä, ja mitkä vähemmän tärkeitä. Ei ole mahdollista antaa yleispäteviä ja kaiken kattavia ohjeita siitä, mitä aspekteja ja niitä edustavia käsitteitä hakuaiheesta pitäisi kulloinkin tunnistaa ja valita kyselyä varten. Erilaiset hakijat voivat päätyä valinnoissaan hyvinkin erilaisiin ratkaisuihin. (Järvelin 1995, 143; 147–148.)

Koska tässä tutkimuksessa käytetään pelkästään alkuperäisestä hakutehtävästä saatavia käsitteitä, hakukäsitteiden tunnistamiseen ja valintaan vaikuttaa erityisen paljon hakupyynnön sanallinen muotoilu. Mitä monisanaisemmin hakupyyntö on ilmaistu, sitä enemmän on hakukäsitteitäkin tarjolla. Lyhyissä ja yksinkertaisissa hakupyynnöissä tärkeimpien pääkäsitteiden tunnistaminen ja valinta on suhteellisen yksinkertaista ja selkeää. Monimutkaisemmissa ja monisanaisemmissa hakupyynnöissä, joissa on useita erilaisia aspekteja ja hierarkiatasoja, käsitteiden valinnasta saattaa muodostua melko subjektiivinen prosessi. Tästä syystä tutkimuksessa kirjataan ylös hakupyynnöt ja niitä edustavat käsitteet sekä toteutetut kyselyt, jotta valintojen pätevyyttä ja tältä osin myös tutkimuksen luotettavuutta on mahdollista arvioida jälkeenpäin.

Toisaalta tässä yhteydessä on hyvä kerrata tämän tutkimuksen päätavoite: missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin. Tutkimuksessa ei siis pyritä etsimään parhaita mahdollisia hakulausekkeita (joilla saadaan paras hakutulos), tai edes hakukoneiden keskivertokäyttäjää parhaiten kuvaavia hakulausekkeita (sillä sellaisten kriteerejä on melko mahdotonta määritellä täysin aukottomasti). Tutkimuksessa ei myöskään vertailla yksittäisten kyselyjen keskinäistä paremmuutta, vaan ainoastaan yritetään ”simulaation” sääntöjen puitteissa tehdä sellaisia hakuja, joilla onnistutaan löytämään relevantteja dokumentteja hakuaiheeseen. Hakulausekkeiden muotoilemisessa käytettävän ”simulaation” ideana onkin tietyssä mielessä vain osoittaa, että siinä esitettyjen ennakkoehtojen (tai sääntöjen) puitteissa on kenen tahansa hakukoneiden keskivertokäyttäjän mahdollista päätyä *vähintään* samaan määrälliseen tulokseen. Tutkimusmenetelmällä saadaan siis aikaiseksi eräänlainen vähimmäistulos. Mahdolliset puutteet käsiteanalyysissä ja/tai hakulausekkeiden muodostamisessa vaikuttavat ainoastaan yhteen suuntaan tutkimustuloksissa: vastauksia voisi löytyä enemmänkin. Mikäli käsiteanalyysissä jää jokin käsite huomaamatta, tai hakulausekkeiden muodostamisessa jää jokin ratkaiseva hakulauseke toteuttamatta, vaikutus tutkimustuloksiin on ainoastaan yhdensuuntainen. Tästä tutkimusasetelmasta seuraa, että tässä tutkimuksessa hakulausekkeiden ei tarvitse olla täydellisen objektiivisia (niitä voisi kutsua subjektiivisiksi valinnoiksi mahdollisimman objektiivisten sääntöjen puitteissa). Tämän takia myöskään käsiteanalyysin ei tarvitse olla objektiivisesti täysin aukoton. Yksittäisten hakupyyntöjen käsiteanalyysien mahdolliset puutteet eivät siten vaikuta tutkimuksen tuloksiin ratkaisevasti (tai vaikutus on ainoastaan yhdensuuntainen).

Hakulausekkeiden konkreettisesta muodostamisesta kerrotaan tarkemmin luvussa 3.7.

3.7 Hakukoneiden keskivertokäyttäjien ”simulaatio”

Arkihavainnot ovat osoittaneet, että ihmisen tiedonhankintaa määrittää useimmiten ns. ”vähimmän vaivan laki”. Sen mukaan ihmiset eivät ole taipuvaisia ponnistelemaan sen enempää kuin tehtävät keskimäärin edellyttävät. Tämä pätee hyvin myös internetin hakukoneiden käytön suhteen. Hakukoneiden käyttäjiä ja heidän syöttämiään kyselyitä on tutkittu lukuisissa erilaisissa tutkimuksissa. Niiden perusteella on voitu todeta, että keskimäärin kyselyt ovat lyhyitä ja hyvin yksinkertaisia, eikä tiedonhakuun haluta käyttää kovinkaan paljoa aikaa. Edellisessä luvussa käytiin läpi aiheeseen liittyviä tutkimuksia ja niiden tuloksia. Näiden tutkimusten perusteella tässä luvussa muodostetaan hakukoneiden keskivertokäyttäjiä ”simuloivia” sääntöjä.

”Simulaation” tarkoituksena on vähentää hakijan yksilöllisten ominaisuuksien, kuten esimerkiksi tiedonhakutaitojen ja mielikuvituksen, vaikutusta hakutuloksiin. Kuten aiemmin jo mainittiin, haun toteuttaminen on aina subjektiivinen prosessi. Yksinkertainenkin hakupyyntö voidaan esittää kyselynä lukuisin eri tavoin. Simulaation sääntöjen avulla voidaan rajoittaa mahdollisten hakujen kokonaismäärää, joten tällä tavoin on mahdollista lisätä tutkimuksen objektiivisuutta ja luotettavuutta. Seuraavaksi käydään läpi näitä ”simulaation” sääntöjä.

Hakutermien määrä

Jotta kyselyistä ei tulisi liian monimutkaisia, on syytä rajata käytettävien hakutermien määrää. Internetin hakukoneisiin syötetyt kyselyt ovat yleensä hyvin lyhyitä. Silversteinin (1999) mukaan 72,4 % Alta Vista –hakukoneeseen lähetetyistä kyselyissä oli yksi- tai kaksisanaisia. Jansen, Spink ja Saracevic (2000) tutkivat Excite-hakukoneeseen lähetettyjä kyselyitä, ja havaitsivat että kyselyt sisälsivät keskimäärin 2,21 termiä. Spinkin ja Jansenin (2004) mukaan hakutermien määrä on lisääntynyt hitaasti kyselyissä, ja lähestyy vähitellen kolmea hakutermiä per kysely. Tässä tutkimuksessa kyselyjen hakutermien määrä rajataan maksimissaan kolmeen hakutermiin. Kaksi hakutermiä on liian vähän (varsinkin monimutkaisemmissa hauissa), ja neljä hakutermiä on jo liikaa, kun ajatellaan keskimääräistä hakukoneiden käyttäjää. Poikkeuksen tähän sääntöön muodostavat kuitenkin erisnimet. Erisnimet (esim. ihmisten, laulujen, elokuvien tms. taideteosten nimet) lasketaan yhdeksi hakukäsitteeksi. Tähän ratkaisuun on päädytty seuraavista syistä:

ensinnäkin moni erisnimi saattaa jo itsessään olla kolme sanaa pitkä, joskus jopa pitempikin. Esimerkiksi kysymys numero 14:

> (08.01.2007) **Kuka esittää Taiskan tunnetuksi tekemän laulun: Moi, moi vaan alkuperäisen version, sen nimi on englanniksi Take me high. Mistä saan kyseisen alkuperäisversion itselleni?**

Mikäli *Take me high* tai *Moi, moi vaan* laskettaisiin kolmeksi hakusanaksi, niin muita hakuja ei voitaisi edes tehdä. Kun edelliset laulun nimet lasketaan yhdeksi hakukäsitteeksi, niihin voidaan yhdistää muita hakutermejä. Tämä on monessa tapauksessa välttämätöntä, mikäli halutaan tehdä järkeviä hakuja. On muistettava, että kolmen hakutermin rajoitus on kuitenkin keinotekoinen, ja se perustuu tutkimuksissa laskettuihin keskiarvoihin. Hakijat tekevät myös yli kolmen hakusanan kyselyitä, varsinkin jos teoksen tms. nimessä itsessään on yli kolme sanaa. Simulaation kannalta on siis perusteltua joustaa säännöstä erisnimien suhteen.

Edellä mainituista syistä myös (lyhyet) sanonnat lasketaan yhdeksi hakukäsitteeksi. Mikäli halutaan tietoa esimerkiksi sanonnan merkityksestä tai etymologiasta, on syytä olettaa että hakukoneen käyttäjä tekisi haun koko sanonnalla, eikä pilkkoisi sanontaa pienempiin osiin.

Edistyneemmät hakutekniikat

Tutkimusten mukaan suuri hakukoneiden käyttäjien enemmistö ei käytä juurikaan edistyneempiä hakutekniikoita. Hölscher (1998) tutki saksalaisen Fireball-tiedonhakujärjestelmän käyttöä. Tutkimuksessa käytiin läpi yli 16 miljoonaa kyselyä. Ainoastaan 3 %:ssa kyselyistä käytettiin Boolen operaattoreita, ja 8 %:ssa kyselyistä käytettiin fraasihakua. Noin 25 % kyselyistä sisälsi ’täytyy sisältää’ –operaattorin (’+’). Silverstein, et al (1999) analysoi lähes miljardi Alta Vista –hakukoneeseen lähetettyä kyselyä. Tutkimuksessa ei eritelty Boolen operaattorien käyttöä, mutta ainoastaan 20,4 % kyselyistä sisälsi joitakin edistyneempiä kyselyoperaattoreita (esim. +. –, & jne). Jansen, Spink ja Saracevic (2000) analysoivat tutkimuksessaan 51 473 Excite-hakukoneen kyselyä. Heidän analyysin mukaan noin 8,5 % kyselyistä sisälsi Boolen operaattoreita, ja noin 9 % kyselyistä sisälsi joitain muita edistyneempiä kyselyoperaattoreita.

Tässä tutkimuksessa ei käytetä edistyneempiä hakutekniikoita, kuten Boolen operaattoreita tai fraasihakuja. Ainoa poikkeus on Boolen AND-operaattori. Se sisältyy tutkimuksessa käytettyyn

Googlen perushakuun oletusarvoisesti. Myös tutkimusasetelman kannalta on perusteltua jättää edistyneemmät hakutekniikat kokonaan pois. ”Simulaation” tarkoituksena on tehdä kyselyjen muodostamisesta mahdollisimman objektiivinen prosessi. Mitä monimutkaisempia kyselyjä on mahdollista muodostaa hakutehtävistä, sitä subjektiivisempia valintoja on myös mahdollista tehdä. Edistyneempien hakutekniikoiden käyttö tekisi kyselyjen muodostamisesta huomattavasti monimutkaisemman prosessin.

Hakutermien valinta

Käytettyjä hakutermejä on myös syytä rajata jollain tavoin. Luonnollisen kielen avulla tapahtuvassa tiedonhaussa on valtavasti erilaisia mahdollisuuksia ilmaista hakutehtävään liittyviä käsitteitä merkkijonotasolla. Jotta hakutermien valinta ei olisi täysin subjektiivista, on syytä pysyä pelkästään alkuperäisissä hakupyynnöissä ilmaistuissa käsitteissä, ja jättää esimerkiksi laajemmat tai suppeammat käsitteet sekä synonyymit kokonaan pois. Kyselyissä käytettävät käsitteet (ja samalla niitä edustavat hakutermit) poimitaan siis pelkästään iGS:ään lähetetyistä kysymyksistä. Tämä tuntuu sinänsä melko luonnolliselta valinnalta. Esimerkiksi Saracevic et al. (tässä Iivonen, 1995, 28) tutkivat valittujen termien alkuperää tiedonhaussa. Koska tutkimuskohteena oli tiedonhaun ammattilaiset, tutkimustulokset eivät kerro hakukoneiden keskivertokäyttäjien hakutottumuksista, vaan tulokset ovat lähinnä suuntaa-antavia. Heidän tutkimuksessaan todettiin että 38 prosenttia tiedonhakuun käytetyistä hakutermeistä oli peräisin suoraan hakupyynnöstä. Todellisuudessa luku voisi olla suurempi, sillä tiedonhakijoilla oli käytettävissään kontrolloitu sanasto. Lisäksi hakua edelsi asiakkaan haastattelutilanne, josta myös saatiin hakutermejä kyselyjä varten.

Vaikka mahdolliset hakukäsitteet rajataan pelkästään iGS-kysymyksestä löytyviin, ei niiden valinta siltikään ole yksinkertainen prosessi. Monet kysymykset ovat varsin monisanaisesti ilmaistuja, ja mahdollisia hakukäsitteitä on runsaasti. Jokainen kysymys analysoidaan yksitellen, sana kerrallaan, ja vain kysymyksen kannalta relevantit käsitteet poimitaan mahdollisiksi hakutermeiksi. Valitut hakukäsitteet merkitään alkuperäisiin iGS-kysymyksiin alleviivattuina. Lisäksi valittuja hakukäsitteitä edustavat hakutermit listataan perusmuodossa (ks. seuraava osio, hakutermien muoto) alakysymyksen alapuolelle arvioituun tärkeysjärjestykseen. Esimerkiksi kysymys numero 110:

(12.01.2007) **Jos tietokoneissa jotain tiedostoja voidaan palauttaa poistamisen jälkeen, niin voidaanko kännyköissä tehdä sama, siis jollain ohjelmalla tai muulla tavalla?**
110: Voiko kännyköiden tiedostoja palauttaa niiden poistamisen jälkeen (jollain ohjelmalla tai muulla tavalla)? (A)
Kännykät, tiedostot, palauttaminen, ohjelma

Hakutermien valintaperiaatteista on kerrottu tarkemmin luvussa 3.6.

Hakutermien muoto

iGS-kysymyksestä löytyneitä hakutermejä ei kuitenkaan käytetä sellaisenaan, koska monissa tapauksissa hakutermit ovat lauserakenteen takia sellaisessa sijamuodossa, ettei siinä muodossa juuri kukaan hakua tekisi. Tässä tutkimuksessa oletetaan, että keskivertokäyttäjä käyttäisi hakutermien perusmuotoa hakuja tehdessään, eikä kysymyksessä esiintyvää sijamuotoa. Tästä syystä valittuja käsitteitä edustavat hakutermit muutetaan nominatiivi- eli perusmuotoon. Perusmuotoistamisessa käytetään apuna Suomea suomeksi – suomen kielen sanakirjaa (1993).

Lisäksi mahdolliset yhdyssanavirheet tai selkeät kirjoitusvirheet korjataan hakutermeistä (tuskin kaikki hakijat tekisivät samoja virheitä). Substantiivien yksikkö- tai monikkomuodoksi valitaan alkuperäisen iGS-kysymyksen muoto. Jos kysymyksessä puhutaan *pyyhkeistä*, hakutermiksi valitaan *pyyhkeet* (ei *pyyhe*).

Substantiivien lisäksi hakukäsitteiksi kelpuutetaan myös verbit. Monissa hakupyynnöissä kysymys on muotoiltu siten, että verbi on välttämätön osa kysymystä, eli se on mahdollinen hakukäsite. Tässä tutkimuksessa oletetaan, että keskivertokäyttäjä ei käyttäisi alkuperäisessä iGS-kysymyksessä esiintyviä verbejä ja niiden sijamuotoja hakusanoina, vaan verbeistä johdettuja substantiiveja eli verbaalisubstantiiveja. Esimerkkinä kysymys numero 1:

(08.01.2007) **Kun elokuvissa lapsi syntyy jossain niin pyydetään tuomaan kuumaa vettä ja pyyhkeitä. Mihin tuota kiehuvaa vettä tarvitaan? Entä pyyhkeitä? Kiehuva vesihän on todella kuumaa...näin olen kuullut.**
1A: Miksi (elokuvissa) synnytyksessä tarvitaan kuumaa vettä? (A) 1B: Miksi (elokuvissa) synnytyksessä tarvitaan pyyhkeitä? (A)
Syntymä, lapsi, kuuma vesi, kiehuva vesi, pyyhkeet, elokuvat

Ilman syntyy (syntyä) –verbiä ei olisi mahdollista muodostaa järkeviä hakulausekkeita, joten se on valittu yhdeksi hakukäsitteeksi. Mutta tuskin monikaan käyttäisi *syntyy* –hakutermiä, tai edes sen perusmuotoa *syntyä*. Verbaalisubstantiivi *syntymä*, tai *syntyminen*, olisivat todennäköisempiä vaihtoehtoja hakutermeiksi.

Toisaalta vaikka tässä tutkimuksessa oletetaan hakijoiden käyttävän hakutermejä tietyssä muodossa, ei taivutusmuotojen valinnalla ole välttämättä aina suurta merkitystä hakutulosten kannalta. Google-hakukone käyttää nimittäin automaattista sanojen stemmausta hyväkseen, joten se ymmärtää hakutermin eri taivutusmuotoja. Hakutermien yksikkö- ja monikkomuotoon sekä sijamuotoon on kuitenkin syytä kiinnittää huomiota, sillä Googlen stemmaus on vielä kaukana täydellisestä. Käytännössä hakutermien muodolla on jonkin verran merkitystä hakutuloksen kannalta.

Kyselyiden muodostaminen

Hakuaiheen käsiteanalyysi –luvussa (luku 3.6) on kerrottu tarkemmin kyselyiden muodostamisesta ja niihin liittyvistä ongelmista. Tässä osassa pyritään selvittämään konkreettisen esimerkin avulla, kuinka hakulausekkeita muodostetaan valituista käsitteistä. Perusperiaate on seuraavanlainen: aluksi käsitteiden joukosta tunnistetaan tärkeimmät pääkäsitteet, eli haun kannalta välttämättömät käsitteet. Pääkäsitteitä edustavat hakutermit esiintyvät kaikissa hakulausekkeissa. Ilman niitä ei järkeviä hakulausekkeita voida muodostaa. Seuraavaksi jäljellä olevat käsitteet arvioidaan ja listataan alakysymyksen alapuolelle tärkeysjärjestykseen, siten että pääkäsite on vasemmanpuoleisin, toiseksi tärkein käsite on seuraavana, ja vähiten tärkein käsite on listan viimeisenä. Kaikkia käsitteitä pyritään käyttämään hakulausekkeita muodostaessa, mutta pääkäsitteet ja tärkeimmät käsitteet ovat etusijalla, eli niitä edustavista hakutermeistä muodostetaan ensimmäiset hakulausekkeet. Yksinkertaisten hakujen suhteen tärkeysjärjestyksellä ei ole kovinkaan suurta merkitystä: kaikkia hakukäsitteitä tullaan todennäköisesti käyttämään hakulausekkeissa. Monimutkaisemmissa hauissa, joissa on runsaasti erilaisia käsitteitä ”tarjolla”, tärkeysjärjestyksen arviointi on erityisen tärkeää, sillä kaikkia mahdollisia hakulausekkeita ei voida toteuttaa.

Esimerkiksi kysymys numero 13:

(08.01.2007) **Kuinka kauan <u>valokuvat</u> <u>säilyy</u> <u>CD-levyllä</u>? On väitetty että vain kymmenen vuotta.**
13: Kuinka kauan valokuvat säilyvät CD-levyllä? (A: ei yksiselitteistä vastausta)
Valokuvat, CD-levy, säilyminen
Valokuvat cd-levy: ei löydy ***Valokuvat cd-levy säilyminen:*** Vastaus löytyy: http://www.yle.fi/pallohallussa/arkistot/300303.html

Yllä olevasta esimerkistä löytyy kolme hakukäsitettä: *valokuvat*, *CD-levy* ja *säilyminen*. *Valokuvat* -käsite on valittu pääkäsitteeksi, koska se on täysin välttämätön kysymyksen kannalta. Sen on oltava mukana kaikissa hakulausekkeissa. Toiseksi tärkein käsite haun kannalta on *CD-levy*. Periaatteessa se voisi olla toinen pääkäsite, mutta se ei ole täysin välttämätön haun kannalta. Tietoa valokuvien säilymisestä CD-levyillä voisi nimittäin löytyä pelkällä *valokuvat säilyminen* -hakulausekkeellakin. Vähiten tärkeimmäksi käsitteeksi on arvioitu tässä tapauksessa käsite *säilyminen*. Se on lähes yhtä tärkeä käsite kuin *CD-levy*, mutta koska kysymyksessä kysyttiin nimenomaan CD-levyillä säilymisestä, *CD-levy* arvioitiin vielä tärkeämmäksi. Edellä mainitut käsitteet on listattu tärkeysjärjestykseen alakysymyksen alapuolelle. Käsitteiden alapuolelle on lueteltu toteutuneet hakulausekkeet ja niiden tulokset.

Hakulausekkeita muodostaessa yhdistellään siis vaihtoehtoisia käsitteitä pääkäsitteeseen (tai pääkäsitteisiin, mikäli niitä on useampia). Edellä mainitussa esimerkissä oli yksi pääkäsite (*valokuvat*) ja kaksi muuta käsitettä (*CD-levy*, *säilyminen*). Näitä käsitteitä voidaan kuvailla kirjaimilla A, B ja C. Kirjaimet edustavat samalla hakulausekkeissa käytettyjä hakutermejä. Esimerkin tapauksessa voitaisiin tehdä seuraavat hakulausekkeet:

A Pelkällä pääkäsitteellä hakeminen voi usein olla järkevä vaihtoehtoehto, mutta tässä tapauksessa kokeiltiin ensin muiden käsitteiden yhdistämistä pääkäsitteeseen.

A + B Pääkäsitteen ja toiseksi tärkeimmän käsitteen yhdistelmä ei tuottanut tulosta.

A + B + C. Oikea vastaus löytyi yhdistämällä kaikki käsitteet. Vastauksen sisältämän nettisivun URL-osoite kirjattiin muistiin.

A + C Pääkäsitteen ja kolmanneksi tärkeimmän käsitteen yhdistelmää ei ehditty kokeilla, sillä oikea vastaus löytyi sitä ennen.

Korkeintaan kolmen hakukäsitteen tapauksissa mahdollisia hakulausekkeita on maksimissaan neljä, eli simulaation sääntöjen puitteissa kaikki voitaisiin tarvittaessa toteuttaa. Mikäli käsitteitä olisi neljä, mahdollisia hakulausekkeita olisi seitsemän: A, A+B, A+C, A+D, A+B+C, A+B+D, A+C+D.

Viiden käsitteen tapauksessa mahdollisia hakulausekkeita olisi jopa yksitoista:
A, A+B, A+C, A+D, A+E, A+B+C, A+B+D, A+B+E, A+C+D, A+C+E, A+D+E

Edellä esitetyt esimerkit osoittavat, että runsaasti hakukäsitteitä sisältävissä hakupyynnöissä mahdollisten hakulausekkeiden määrä kasvaa nopeasti niin suureksi, ettei kaikkia mahdollisia hakulausekkeita ole järkevää toteuttaa. Kyselyjen määrää on siis rajattava jollain tavoin. Tästä seuraa, että käsitteiden tärkeysjärjestyksen arvioinnilla, pääkäsitteiden valinnalla sekä hakulausekkeissa käytettävien hakutermien valinnalla on melko suuri merkitys hakutulosten kannalta. Monimutkaisimmissa hauissa subjektiivisten valintojen mahdollisuus siis kasvaa jonkin verran. Kuten luvussa 3.6 mainittiin, tällä ei ole ratkaisevaa merkitystä tutkimustulosten kannalta (vaikutus on ainoastaan yhdensuuntainen).

Kyselyjen määrä

Toteutettavien kyselyjen määrää on myös rajattava tietyissä tapauksissa. Mikäli hakupyynnössä esiintyy useampia haun kannalta sopivia käsitteitä, on edellä mainituista rajauksista huolimatta mahdollista muodostaa lukuisia erilaisia kyselyjä. Tässä tutkimuksessa haun onnistumisen kriteerinä on yksikin hakutuloksessa esiintyvä relevantti dokumentti (ks. luku 3.8). Mikäli relevantti dokumentti löytyy esimerkiksi jo ensimmäisellä kyselyllä, ei muita mahdollisia kyselyjä tarvitse enää toteuttaa kyseisen hakupyynnön osalta. Toisaalta ei ole tarkoituksenmukaista tehdä hakuja jokaisella mahdollisella hakulausekkeella – eihän keskivertohakijakaan varmasti niin tekisi. Valitettavasti on melko hankalaa määritellä jotain tiettyä kyselyjen määrää, johon keskivertokäyttäjä myös päätyisi. Vaikka joissakin tutkimuksissa on laskettu hakijoiden keskimääräisiä kyselyjen määriä yhtä hakutehtävää kohden (esimerkiksi Jansen et al. 2000: 2,8 kyselyä), kyselyjen määrä on riippuvainen monista tekijöistä, kuten hakijasta sekä hakuaiheesta ja sen vaativuudesta. Ei voida sanoa, että keskivertohakija tekisi aina maksimissaan X kyselyä. Tutkimuksen kannalta on kuitenkin syytä rajata hakujen määrää jollain tavoin. Tässä tutkimuksessa

hakuja on tehty 1–3 kappaletta per alakysymys. Joissakin tapauksissa hakuja on tehty enemmän (maksimissaan viisi).

Dokumenttien kieli

Koska otoksen iGS-kysymykset ovat suomenkielisiä, tehdään myös kyselyt suomen kielellä. Tämä rajaa myös löytyneet dokumentit pääasiassa suomenkielisiin. Joukkoon saattaa kuitenkin eksyä myös esimerkiksi englanninkielisiä dokumentteja, mikäli haetaan esimerkiksi pelkästään erisnimillä. Joillekin hakijoille tällaiset dokumentit saattaisivat olla hyvinkin relevantteja, toisille taas täysin hyödyttömiä, koska kielitaito ei riitä dokumentin lukemiseen. Koska tarkoitus on ”simuloida” hakukoneiden keskivertokäyttäjää, emme voi tietää hakijan kielitaidoista juuri mitään. Tästä syystä vieraskieliset dokumentit sivuutetaan hakutuloksista kokonaan.

Tulossivujen katsominen

Koska tässä tutkimuksessa tehdään relevanssiarvio hakutuloksien dokumenteille, eikä kaikkia mahdollisia hakutuloksen dokumentteja voida käydä läpi, on arvioitavien dokumenttien määrää rajattava jollain tavoin. Tutkimusten mukaan (mm. Jansen, Spink & Saracevic 2000; Silverstein et al. 1999) useimmat hakukoneiden käyttäjät käyvät läpi ainoastaan ensimmäisen hakutulossivun dokumentit eli 10 ensimmäistä dokumenttia. Tätä 10 dokumentin rajaa on käytetty muissakin tutkimuksissa (mm. Lucas & Topi 2002), joten sitä voidaan pitää sopivana rajana tässäkin tutkimuksessa.

Google-hakutuloksesta avautuvia dokumentteja tarkastellaan pelkästään näkyviltä osin. Relevantin vastauksen täytyy siis löytyä suoraan dokumentista, ilman linkkien avaamista. Relevanssiarviosta kerrotaan tarkemmin seuraavassa luvussa.

3.8 Löydettyjen dokumenttien relevanssiarvio

Tiedonhaun tavoitteena on löytää hakijalle relevanttia informaatiota, useimmiten relevanttien dokumenttien muodossa. Jotta haun onnistumista voitaisiin mitata, pitää siis määritellä, mitkä

tulosjoukon dokumenteista ovat relevantteja eli oleellisia. Relevanssilla tarkoitetaan tiedonhakijan tiedontarpeiden ja löytyneiden dokumenttien käyttökelpoisuuden vastaavuutta. Relevanssin käsite voidaan jakaa kahteen alakäsitteeseen: aiherelevanssiin ja käyttäjärelevanssiin. Aiherelevanssilla tarkoitetaan sitä, että dokumentti käsittelee hakupyynnön määrittelemää aihetta, eli relevanssiarviossa keskitytään hakukysymyksen ja dokumenttien kuvausten väliseen täsmäävyyteen. Käyttäjärelevanssi huomioi dokumentin aiheen lisäksi tiedon käyttäjän arvion dokumentin käyttökelpoisuudesta. Käyttäjän arvioon vaikuttavat monet asiat, kuten tiedontarpeen aiheuttavan tehtävän luonne, dokumentin kieli, ulkoasu sekä käyttäjän aikaisemmat tiedot kyseisestä aiheesta. (Järvelin & Sormunen 1999, 117; Alaterä & Halttunen 2002, 125–126.)

Koska iGS-verkkotietopalvelun asiakkaiden varsinaisesta tiedontarpeesta ei ole muuta vihjettä kuin iGS:ään esitetty kysymys, löydettyjä dokumentteja arvioidaan tässä tutkimuksessa pääasiassa aiherelevanssin näkökulmasta. Dokumentteja voidaan arvioida tietyssä mielessä myös käyttäjärelevanssin kannalta, vaikkei tiedon käyttäjien arviota vastausten käyttökelpoisuudesta olekaan saatavilla. iGS-palvelun vastaajat ovat nimittäin vastausta hakiessaan ja muodostaessaan tehneet oman arvionsa asiakkaan tiedontarpeesta ja dokumenttien käyttökelpoisuudesta. Tätä vastauksissa ilmenevää tiedonhaun ammattilaisen muodostamaa arviota voidaan pitää melko käyttökelpoisena mittarina ja vertailukohtana dokumenttien relevanttiutta arvioitaessa. Löytyneiden dokumenttien relevanttiutta arvioitaessa voidaan siis aiherelevanssin lisäksi ottaa huomioon myös kysymykseen annettu alkuperäinen vastaus lähteineen. Käytännössä kysymykset, vastaukset sekä niissä käytetyt lähteet muodostavat yhdessä kriteerit relevanssiarviolle.

Koska verkossa periaatteessa kuka tahansa voi julkaista melkein mitä tahansa, lähdeaineiston luotettavuuden arviointi on normaalissa tiedonhaussa keskeisessä osassa. Tässä tutkimuksessa lähteiden luotettavuuden arvioinnille ei anneta yhtä suurta painoarvoa. Relevanssiarviota tehtäessä dokumentteja verrataan alkuperäisiin iGS-vastauksiin, joten pääpaino on alkuperäisen iGS-vastauksen mukaisen vastauksen löytymisessä. Tutkimuksessa oletetaan, että iGS-vastaaja on löytänyt kysymykseen oikean vastauksen, ja lisäksi tarkastanut vastauksen oikeellisuuden luotettavasta lähteestä. Tästä syystä sellaisetkin lähteet, joita normaalisti ei välttämättä kelpuutettaisi luotettavien lähteiden joukkoon, hyväksytään vastauksen lähteeksi, mikäli lähteestä löytynyt informaatio vain on iGS-vastauksen mukaista. Tällaisia ns. epäluotettavia lähteitä ovat esimerkiksi Wikipedia ja keskustelupalstat. Toisaalta myös iGS-vastaajat käyttävät joissakin vastauksissaan tällaisia lähteitä. Tässäkin mielessä ne ovat kelvollisia lähteitä tämän tutkimuksen

kannalta. Tässä tutkimuksessa ainoastaan blogeja ei kelpuuteta vastauksen lähteeksi. Blogeja ei ole ensisijaisesti tarkoitettukaan luotettavan tiedon lähteiksi, vaan enemmänkin mielipiteiden julkaisemiskanavaksi.

Tässä tutkimuksessa relevanssi nähdään binäärisenä ominaisuutena: löydetty dokumentti on joko relevantti tai epärelevantti. Vaikka relevantit dokumentit voitaisiin jaotella asteittain esimerkiksi marginaalisesti relevantteihin, relevantteihin ja erittäin relevantteihin dokumentteihin, tässä tutkimuksessa tämä ei ole tarpeen. Tämän tutkimuksen tavoitteenahan on selvittää, että missä määrin Google-hakukoneella on mahdollista löytää relevantteja vastauksia iGS-palvelun kysymyksiin. Haun onnistumisen kriteeriksi riittää siis se, että löytyykö hakutuloksesta yhtään relevanttia dokumenttia. Marginaalisesti relevanteilla dokumenteilla ei ole mitään merkitystä hakupyynnön esittäjälle, ja osittain relevantit dokumentit eivät myöskään ole riittäviä haun onnistumisen kannalta. Koska hakutuloksien dokumenttien arvioinnissa riittää että löydetään yksi relevantti dokumentti, relevanssiarviota ei tarvitse tehdä kaikkien löydettyjen dokumenttien osalta. Kun yksi relevantti dokumentti on löytynyt, muita dokumentteja ei tarvitse enää arvioida. Tämän löytyneen dokumentin URL-osoite kirjataan muistiin, jotta dokumentin relevanssiarvion pätevyys voidaan tarkastaa jälkeenpäin.

3.9 Aineiston analysointi

Tutkimuksen aineisto käsiteltiin tilastollisesti SPSS-ohjelmaa apuna käyttäen. Google-hakujen ja relevanssiarvioiden jälkeen alakysymykset, niiden kategoriat sekä kysymyksiin liittyvät hakutulokset (löytyi / ei löytynyt) syötettiin SPSS-ohjelmaan. Tämän jälkeen ohjelmalla luotiin aineistoa kuvaavia taulukoita, sekä ristiintaulukoinnin avulla tarkasteltiin muuttujien (kysymysten kategoriat) yhteyttä Google-hakukoneen avulla löytyneisiin hakutuloksiin. Muuttujien välistä tilastollista yhteyttä tutkittiin Pearsonin Khi-2 –testin avulla. Tulokset raportoidaan sanallisessa muodossa ja taulukoina luvussa 4.

4. TUTKIMUSTULOKSET

Tässä luvussa raportoidaan empiirisen tutkimuksen tulokset. Seuraavissa alaluvuissa tullaan tarkastelemaan mm. relevanttien suomenkielisten vastausten määrää sekä otoksen kysymysten jakautumista eri kysymystyyppeihin ja iGS-palvelun kategorioihin. Lisäksi tarkastellaan Google-hakujen tuloksia kysymystyypeittäin.

4.1 Relevanttien suomenkielisten vastausten määrä

”Simulaation” avulla Google-hakukoneella löytyi relevantti suomenkielinen vastaus 188 kysymykseen. Otoksen 509 kysymyksestä tämä oli 36,9 prosenttia, eli runsas kolmasosa. 321 kysymykseen (63,1 prosenttia) ei vastausta löytynyt Googlella.

Tämän tutkimuksen mukaan Helsingin kaupunginkirjaston iGS-verkkotietopalveluun lähetetyistä kysymyksistä runsas kolmasosa oli siis sellaisia, joihin periaatteessa kuka tahansa hakukoneiden keskivertokäyttäjä voisi löytää ”simulaation” sääntöjen puitteissa vastauksen (eli käytännössä suhteellisen helposti). Toki suurimpaan osaan kysymyksistä ei löytynyt vastausta Googlella, mutta silti lähes 37 prosentin osuus tuntuu varsin suurelta. Tutkimuksen ”simulaation” säännöissä rajoitettiin nimittäin melko lailla hakutermien valintaa ja hakulausekkeiden muodostamista. Millainen olisi ollut tulos, mikäli Google-hakuja ei olisi rajattu näin paljon?

4.2 Kysymysten jakautuminen kysymystyyppeihin ja iGS-kategorioihin

Otoksen kysymykset jaettiin kolmeen kysymyskategoriaan. Faktakysymyksiä ja tiettyä aihetta koskevia kysymyksiä oli lähes yhtä paljon: Tiettyä aihetta koskevia kysymyksiä oli eniten (225 kpl eli 44,2 prosenttia). Toiseksi eniten oli faktakysymyksiä (222 kpl eli 43,6 prosenttia). Tiettyä teosta koskevia kysymyksiä oli selvästi vähiten (61 kpl eli 12,2 prosenttia). Kysymystyyppien jakauma on esitetty taulukossa 1.

Taulukko 1: iGS-kysymykset kysymystyypeittäin (n = 509).

Kysymyksen kategoria	**n**	**%**
Tiettyä aihetta koskevat kysymykset	225	44,2
Faktakysymykset	222	43,6
Tiettyä teosta koskevat kysymykset	62	12,2
Kaikki kysymykset yhteensä	509	100

Kysymystyyppien jakauma on selvästi erilainen verrattuna Gräsbeckin (2008, 28) saamiin tuloksiin. Kysy kirjastonhoitajalta –palveluun lähetettiin eniten tiettyä teosta koskevia kysymyksiä (33 prosenttia), ja toiseksi eniten tiettyä aihetta koskevia kysymyksiä (25 prosenttia). Faktakysymyksiä oli Gräsbeckin tutkimuksessa kolmanneksi eniten (23 prosenttia). Neljänneksi eniten Gräsbeckin tutkimuksessa tuli kirjastonkäyttöä/kirjaston toimintatapoja koskevia kysymyksiä (15 prosenttia). Vastaavanlaisia (HelMet-aineistohakuun liittyviä) kysymyksiä tuli tässä tutkimuksessa vain pari kappaletta.

Kun tarkastellaan kysymystyyppejä sen mukaan löytyikö vastaus Googlella, taulukosta 2 voidaan havaita että tämä onnistui parhaiten faktakysymysten osalta. Hieman yli 41 prosenttiin (92 kpl) faktakysymyksistä löytyi vastaus Googlella. Tiettyä aihetta koskeviin kysymyksiin löytyi vastaus kolmasosaan (33,8 prosenttia) kysymyksistä. Tiettyä teosta koskeviin kysymyksiin löytyi vastaus huonoiten: alle kolmasosaan (30,5 prosenttia). (ks. taulukko 2).

Taulukko 2. Vastausten löytyminen Googlella kysymystyypeittäin (n = 509)

Vastauksen löytyminen Googlella	**Kysymyksen tyyppi**			
	Fakta	Aihe	Teos	Yhteensä
Vastaus löytyi Googlella	92 (41,4 %)	76 (33,8 %)	20 (32,3 %)	188 (36,9 %)
Vastausta ei löytynyt Googlella	130 (58,6 %)	149 (66,2 %)	42 (67,7 %)	321 (63,1 %)
Yhteensä	222 (100 %)	225 (100 %)	62 (100 %)	509 (100 %)

Pearsonin Khi-2 testin mukaan muuttujien välillä ei kuitenkaan ollut tilastollista merkitsevää yhteyttä ($p = 0.175$). Kysymysten tyypeillä ei siis ollut tilastollisesti merkitsevää yhteyttä siihen,

että missä määrin Google-hakukoneella löytyy tai ei löydy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin.

iGS-kysymysten analyysissä käytettiin myös iGS-vastaajien tekemiä luokitteluja. Vastaajat luokittelevat kysymyksen aiheen vastaamisen yhteydessä, iGS-vastauslomakkeesta löytyvien valmiiden kategorioiden mukaisesti. Vastaajat voivat luokitella kysymyksen useampaan luokkaan, mutta useimmiten kysymyksille oli valittu vain yksi kategoria. Mikäli kysymyksellä oli useampia luokkia, pyrittiin tässä tutkimuksessa valitsemaan näistä parhaiten kysymystä kuvaava luokka.

iGS:n vastauslomakkeessa kategorioita on yhteensä 62. Koska kolme kategoriaa eivät liittyneet kysymyksen aiheeseen, niitä ei huomioitu analyysissä. Nämä kategoriat ovat:

- Lasten kysymykset (kysymys on lähetetty Helsingin kaupunginkirjaston lastensivujen puolelta),
- Radiokysymykset (kysymys on esitetty Ylen aikaisen radiolähetyksessä) ja
- Ylen aikainen –kysymykset (kysymys on tullut Ylen aikaisen verkkosivujen kautta).

Kysymysten jakautuminen iGS-kategorioiden mukaisesti yksilöidään taulukossa 3.

Otoksen kysymykset jakautuivat verrattain tasaisesti eri aihepiireihin. Ainoastaan kielet, kielitiede, sanat –kategoria erottui selvemmin muista (13,2 prosenttia). Musiikkiaiheisia kysymyksiä esiintyi toiseksi eniten (7,3 prosenttia), ja lähes yhtä paljon esitettiin terveydenhoitoon ja lääketieteeseen liittyviä kysymyksiä (7,1 prosenttia). Seuraavaksi eniten oli historia-aiheisia kysymyksiä (5,3 prosenttia). Muut kategoriat olivat alle neljän prosentin kokoluokkaa.

Karisen (2008, 47) tutkimuksessa oli hieman tasaisempi jakauma, mutta kärkikaksikko oli sama. Kielet, kielitiede, sanat –kategoriaan tuli eniten kysymyksiä (9,9 prosenttia), ja toiseksi eniten tuli musiikkiaiheisia kysymyksiä (5,5 prosenttia). Seuraavaksi suurimmat kategoriat olivat kirjailijat, kirjallisuus (4,9 prosenttia) ja ruoka ja juomat (4,9 prosenttia). Karisen tutkimuksessa historia-kategoria oli vasta sijalla 12. (3 prosenttia).

Taulukko 3. Kysymysten iGS-kategorioiden mukainen luokittelu (n = 509)

Luokka	kpl	%	Luokka	kpl	%
Kielet, kielitiede, sanat	67	13,2	Uskonnot	6	1,2
Musiikki	37	7,3	Perinteet ja tapahtumat	5	1,0
Terveydenhoito, lääketiede	36	7,1	Tekniikka	5	1,0
Historia	27	5,3	Elektroniikka	4	0,8
Ruoka ja juomat	19	3,7	Internet, sähköposti	3	0,6
Elokuva, video, teatteri, tanssi	18	3,5	Kalastus, veneily, metsästys ym.	3	0,6
Opiskelu, kasvatus	17	3,3	Matematiikka, mitat, suureet	3	0,6
Ajoneuvot, kulkuneuvot, liikenne	16	3,1	Muoti, tyyli, pukeutuminen	3	0,6
Luonnontieteet, luonto	14	2,8	Rakentaminen	3	0,6
Maantiede, maat, paikat	14	2,8	Taiteet	3	0,6
Tietotekniikka	14	2,8	Ennätykset	2	0,4
Julkishallinto ja viranomaiset	13	2,6	Erotiikka	2	0,4
TV, radio, lehdet	13	2,6	iGS	2	0,4
Koti, asuminen	12	2,4	Kauneus, kauneudenhoito	2	0,4
Kasvit, puutarha	11	2,2	Matkustaminen, matkailu	2	0,4
Kirjailijat, kirjallisuus	10	2,0	Tilastot	2	0,4
Talous, yritykset, kuluttaminen	10	2,0	Ihmissuhteet	1	0,2
Urheilu, liikunta	10	2,0	Kirjastot	1	0,2
Yhteiskunta, politiikka	10	2,0	Leikit ja pelit	1	0,2
Ympäristö, ympäristönsuojelu	10	2,0	Sukututkimus	1	0,2
Käden taidot, korjaaminen	9	1,8	Teollisuus	1	0,2
Eläimet ja lemmikit	8	1,6	Trivia	1	0,2
Erikoiset	8	1,6	Arkkitehtuuri	0	0
Harrastukset, keräily	8	1,6	Filosofia	0	0
Ihmiset	8	1,6	Kansat, kulttuurit	0	0
Työ, työelämä	8	1,6	Matkapuhelimet, puhelimet	0	0
Helsinki	7	1,4	Rajatieteet, mystiikka	0	0
Lait, lainsäädäntö	7	1,4	Sarjakuvat	0	0
Avaruus, tähtitiede, aika	6	1,2	Sää, ilmasto	0	0
Sota, armeija	6	1,2	**Yhteensä**	509	100

Neljää laajinta kategoriaa tarkasteltiin myös kysymystyypeittäin. iGS-kategorioiden jakautuminen kysymystyypeittäin on esitetty taulukossa 4.

Taulukko 4. Neljän suurimman iGS-kategorian jakautuminen kysymystyypeittäin (n = 167)

Luokka	Fakta		Aihe		Teos		Yhteensä	
	kpl	%	kpl	%	kpl	%	kpl	%
Kieli, kielitiede, sanat	43	64,2	24	35,8	0	0,0	67	100
Musiikki	6	16,2	6	16,2	25	67,6	37	100
Terveydenhoito, lääketiede	14	38,9	22	61,1	0	0,0	36	100
Historia	13	48,1	14	51,9	0	0,0	27	100

Kieli, kielitiede ja sanat –kategoriassa lähes kaksi kolmasosaa (64,2 prosenttia) oli faktakysymyksiä. Tässä kategoriassa noin kolmasosa (35,8 prosenttia) oli tiettyä aihetta koskevia kysymyksiä. Tämän kategorian kysymyksissä tiedusteltiin mm. sanojen alkuperää tai sanojen iimerkityksiä.

Esimerkki 1

Mistä sanat asiakirja ja arkisto ovat peräisin? Latinan- vai kreikankielestä vai mistä?

Neljästä suurimmasta iGS-kategoriasta musiikki oli ainoa missä esiintyi tiettyä teosta koskevia kysymyksiä. Musiikkiaiheisista kysymyksistä suurin osa olikin tiettyä teosta koskevia kysymyksiä (67,6 prosenttia). Kysymyksissä tiedusteltiin mm. kappaleiden esittäjiä tai kappaleiden nimiä, tai vaikkapa tietoa jostain yhtyeestä.

Esimerkki 2

1980-luvun taitteesta muistan rekkamies-aiheisen kappaleen, jossa laulettiin jotain ”teiden supermiehestä”. Esittäjä oli mies. Löytäisittekö esittäjän ja kappaleen nimen? Kappale oli enemmänkin humoristinen kuin Matti Eskon Rekkamies –tyylinen.

Esimerkki 3

Mistä löytyy tietoa poikabändi McFly:sta??

Terveydenhoito ja lääketiede –kategoriassa oli eniten tiettyä aihetta koskevia kysymyksiä (61,1 prosenttia). Kysymykset oli esitetty useimmiten melko yleisellä tasolla. Muutamat kysymykset olivat hyvinkin henkilökohtaisia, ja ne liittyivät mm. terveysasioihin.

Esimerkki 4

Kärsin niin kovasta iskiaskivusta, että näkö sumenee enkä pysty lähtemään ulos. Selkäni on viipalekuvattu ja hermoratatutkimus tehty. 4-5 lannenikamien välissä on välilevyn pullistuma. Kirurgisen sairaalan kirurgi suhtautui minuun kuin häiritsevään kärpäseen eikä allekirjoittanut mitään hoitoa. Terveysasemalta ei osattu neuvoa, mitä pitäisi tehdä. Tiedän että hoitoa saisi yksityissairaalasta rahalla, mutta olen työtön, enkä pysty maksamaan sen sektorin leikkausta enkä muuta hoitoa. Minkä tahon puoleen minun pitäisi kääntyä saadakseni asianmukaista hoitoa?

Historia-aiheiset kysymykset jakautuivat hyvin tasaisesti fakta- ja aihekysymysten kesken. Tämän ryhmän kysymykset olivat toisinaan hieman hankalia luokittelun kannalta: historia-kategorian lisäksi kysymyksen olisi voinut usein luokitella johonkin toiseen kategoriaan.

Esimerkki 5.

> *Isoäitini kirjoittaa kirjeessä 23.6.1919: ”Eilen kävin kirkossa ja kävin syömässä automaatissa. Sitten menin kotiin syömään piimää.” Mikä tuo automaatti voisi olla? Kirkko on Johanneksen kirkko Helsingissä ja koti on Neitsytpolulla, joten ”automaatti” sijaitsee jossain niillä paikkeilla.*

Tämän kysymyksen olisi periaatteessa voinut luokitella historia-kategorian sijaan/lisäksi Helsinki-kategoriaan, koska kysymys liittyi Helsinkiin. Kysymys luokiteltiin kuitenkin historia-aiheiseksi kysymykseksi, koska tapahtumasta oli kulunut jo varsin pitkä aika. Helsinki-kategoriaan valitut kysymykset liittyivät joko nykyhetkeen tai lähimenneisyyteen.

Seuraavissa alaluvuissa tarkastellaan iGS-kysymysten kysymystyyppejä. Niitä tarkastellaan Google-hakutuloksiin nähden. Esimerkkien avulla pyritään havainnollistamaan, minkätyyppisiin kysymyksiin löytyi vastaus Googlella, ja millaisiin ei löytynyt. Kysymystyyppien analyysissä tarkastellaan myös iGS-kategorioiden jakautumista. Koska erilaisia iGS-kategorioita on varsin runsaasti (59 kpl), ja kysymykset jakautuivat melko tasaisesti eri kategorioihin, yksittäisten kategorioiden kysymysten lukumäärät jäivät varsin pieniksi. Tästä syystä yksittäisistä iGS-kategorioista ei voida tehdä kuin suuntaa-antavia havaintoja Google-hakutulosten suhteen.

4.3 Faktakysymykset

Faktakysymyksistä lähes viidesosa (19,4 prosenttia) kysymyksistä liittyi iGS-kategoriaan kieli, kielitiede, sanat. Toiseksi eniten oli terveydenhoitoon ja lääketieteeseen liittyviä kysymyksiä (6,3 prosenttia), mutta lähes yhtä paljon oli historiaan (5,9 prosenttia) tai ajoneuvoihin, kulkuneuvoihin ja liikenteeseen liittyviä kysymyksiä (4,5 prosenttia). Suuria prosentuaalisia eroja eri kategorioiden välillä ei löytynyt (lukuun ottamatta kategoriaa kieli, kielitiede, sanat).

Kysymykset joihin löytyi vastaus Googlella

Kategoriaan terveydenhoito, lääketiede löytyi parhaiten vastaus Googlella. Lähes kahteen kolmasosaan faktakysymyksistä löytyi vastaus (64,3 prosenttia).

Esimerkki 6

> *Olen etsinyt tietoa siitä, kauanko ruoan matka läpi ruuansulatuskanavan ja koko kehon kestää? Jotkut sanovat 24 tuntia ja jotkut paljon vähemmän. Mistä kaikesta se on riippuvaista, siis vaikuttaako nautittu ravinto? Onko esim. lihalla ja vaikkapa salaatinlehdellä eri ”läpikulkunopeus”?*

Esimerkin kysymys voidaan muotoilla ja jakaa kahteen alakysymykseen:

- Kuinka kauan ruoalla kestää kulkeutua ruoansulatuskanavan läpi?
- Miten nautittu ravinto vaikuttaa ruoan kulkunopeuteen (ruoansulatuskanavan läpi)?

Molempiin kysymyksiin löytyi vastaus hakusanalla *ruoansulatuskanava*. Hakusanalla löytyi Finfoodin nettisivu, missä kerrottiin ruoan matkasta elimistössä (3-4 vuorokautta), sekä mm. tietoa hiilihydraatti- ja proteiinipitoisten ruokien sekä kuitujen vaikutuksesta ruuan läpimenoaikaan. Samaa lähdettä oli käytetty alkuperäisessä iGS-vastauksessakin.

Googlella löytyi vastaus melko hyvin myös kieli, kielitiede, sanat –kategorian kysymyksiin (48,8 prosenttia). Useimmiten kysyttiin joko sanojen etymologiaa tai sanojen merkitystä.

Esimerkki 7

> *Itä-Helsingissä on Humikkalantie, Untamalantie ja Lupajantie. Humikkala lienee jonkinlainen muinainen hautapaikka. Mutta mitä ovat Untamala ja varsinkin Lupaja??*

Tämäkin kysymys täytyi myös jakaa kahteen alakysymykseen: Mikä on Untamala, ja mikä on Lupaja? Hakulausekkeella *Lupaja hautapaikka* löytyi vastaus molempiin kysymyksiin: Museoviraston Hoidetut muinaisjäännöskohteet vuodelta 2004 –nettisivulta löytyi luettelo kohteista selityksineen (Untamala: rautakautinen asuinpaikka ja kalmistoalue Laitilassa. Lupaja: rautakautinen kalmisto Perniössä).

Kysymykset joihin ei löytynyt vastausta Googlella

Historia-aiheiset kysymykset osoittautuivat ongelmallisiksi Google-hakujen kannalta: ainoastaan 15,4 prosenttiin historia-aiheisista kysymyksistä löytyi vastaus Googlella. Faktakysymykset historiasta vaihtelivat aiheeltaan laidasta laitaan, maailman ensimmäisestä keinutuolista aina vuoden

1979 äitiyspakkauksen sisältöluetteloon. Fakta-kysymysten vastaukset eivät aina olleet yksittäisiä faktoja, vaan vastaukset saattoivat olla myös pitkiä luetteloitakin (kuten äitiyspakkauksen sisältöluettelo). Monissa iGS-vastauksissa käytettiin historia-aiheisten kysymysten vastauksissa kirjallisia lähteitä. Ehkäpä se omalta osaltaan selittää Google-hakujen heikkoa tulosta historian osalta.

Esimerkki 8

Kuinka paljon asukkaita oli Turussa vuonna 1603?

Hakulausekkeet (missä yhdistelin hakusanoja *Turku*, (*vuosi) 1603*, *asukkaat*) eivät tuottaneet tulosta. Käytettävissä olevat hakusanat eivät olleet riittävän täsmällisiä historia-aiheisten sivujen etsimiseen. Lisäksi luku 1603 saattoi esiintyä esim. puhelinnumeroissa. Toisaalta vastausta ei luultavasti löytyisi netistä lainkaan. Alkuperäisessä vastauksessa vastaaja nimittäin itse antoi vain arvioidun asukasluvun (oma tietämys), ja kertoi samalla ettei tarkkoja lukumääriä tiedetä, koska väestönlaskenta ei ollut kehittynyttä.

Toinen laajempi kategoria mikä ei tuottanut tulosta Google-hakujen kannalta, oli ajoneuvot, kulkuneuvot, liikenne (vain joka viidenteen kysymykseen löytyi vastaus).

Esimerkki 9

Miksi Finnairin lentojen tunnus on AY? Mistä se tulee?

Kysymyksiin ei löytynyt vastausta hakusanojen *Finnair, lennot, tunnus, AY* –yhdistelmillä. Monilla nettisivuilla kerrottiin Finnairin lentojen tunnusten olevan AY, mutta selitystä kirjaimille ei löytynyt. Vastaus oli tosin lähellä löytymistä, sillä *Finnair*-hakusanalla löytyi Wikipedian artikkeli, missä olisi ollut linkki IATA-artikkeliin, mistä olisi löytynyt vastaus kysymykseen. Tutkimuksen ”simulaation” sääntöjen mukaan vastaus pitää kuitenkin löytyä suoraan Google-hakutuloksen sivulta, joten vastausta ei siis löytynyt.

4.4 Tiettyä aihetta koskevat kysymykset

Tiettyä aihetta koskevista kysymyksistä eniten kysymyksiä kohdistettiin kategoriaan kielet, kielitiede, sanat (24 kpl eli 10,7 prosenttia). Terveydenhoito ja lääketiede –aiheisia kysymyksiä tuli

toiseksi eniten (22 kpl eli 9,8 prosenttia). Seuraavaksi suurimmat iGS-kategoriat olivat historia (6,2 prosenttia), ruoka ja juomat (5,3 prosenttia), koti, asuminen (4,9 prosenttia) sekä opiskelu, kasvatus (4,9 prosenttia). Koti, asuminen –kategorian kysymykset liittyivät selvimmin yksittäisiin aiheisiin: otoksen 11:sta kysymyksestä 10 koski tiettyä aihetta, ja vain yksi oli faktakysymys.

Kysymykset joihin löytyi vastaus Googlella

Edellä mainituista iGS-kategorioista oli Google-hakujen kannalta tuloksellisin terveydenhoito, lääketiede: Googlella löytyi vastaus noin 59 prosenttiin kysymyksistä. Kysymykset liittyivät usein joko omaan terveyteen (esim. aiheena haiskahtava päänahka), tai sitten kysyjät olivat vain uteliaita tietämään jostain aiheesta (esim. miksi elokuvissa käytetään kuumaa vettä ja pyyhkeitä synnytyksessä).

Esimerkki 10

Vaikuttaako useiden eri alkoholijuomien nauttiminen krapulan voimakkuuteen?

Hakulausekkeella *krapula voimakkuus* löytyi päihteistä ja riippuvuuksista kertova Päihdelinkki –sivusto, missä kerrottiin perusteellisesti erilaisten alkoholijuomien ja krapulan yhteyksistä toisiinsa.

Esimerkki 11

Jos tietokoneissa jotain tiedostoja voidaan palauttaa poistamisen jälkeen, niin voidaanko kännyköissä tehdä sama, siis jollain ohjelmalla tai muulla tavalla?

Hakulausekkeella *Kännykät tiedostot palauttaminen* löytyi Ontrack EasyRecovery –ohjelmiston nettisivut. Ohjelmiston avulla voi palauttaa tiedostoja myös kännyköiden muistikorteilta. Lisäksi hakulausekkeella *tiedostot palauttaminen* löytyi erään tiedonpalautusyrityksen (Gelkin Tmi) nettisivut. Kyseinen yritys palauttaa tiedostoja mm. GSM-puhelimien muistikorteilta ja sisäänrakennetusta muistista.

Kysymykset joihin ei löytynyt vastausta Googlella

Historia-aiheiset kysymykset osoittautuivat myös tiettyä aihetta koskevien kysymysten kohdalla ongelmallisiksi: ryhmän 14:sta kysymyksestä ainoastaan yhteen löytyi vastaus Googlella. Kysymykset vaihtelivat aiheeltaan hyvin paljon: jossain kysymyksessä tiedusteltiin mm. porvarivalan sisältöä, toisessa taas haluttiin tietää mahdollisimman paljon luovutetun Karjalan linnavuorista.

Historia-aiheiset kysymykset olivat luokittelun kannalta usein ongelmallisia. Historiaa käsittelevät kysymykset liittyivät hyvin usein johonkin toiseen kategoriaan: esim. Helsingin historiaa käsittelevät kysymykset olisi voinut luokitella Helsinki-kategoriaan. Samoin sotahistoriaa koskevat kysymykset olisi voinut luokitella Sota, armeija –kategoriaan. Jopa terveydenhoito, lääketiede –kategorian kysymykset saattoivat olla joskus historiaa sivuavia. Johdonmukaisuuden vuoksi päätin kuitenkin luokitella kaikki tällaiset tapaukset historia –kategoriaan.

Esimerkki 12

> *Kysyn vielä toisenkin kysymyksen.. Muistan kuulleeni jossain ”astmatupakasta” jota lääkärit määräsivät astmaa sairastaville poltettavaksi. Kuulostaa erikoiselta, onko sellaista joskus apteekista saanut? Koska? Mistähän aiheesta löytyisi tietoa?*

Kysymyksestä poimituilla hakusanoilla *astmatupakka*, *astma*, *apteekki* ei löytynyt relevanttia vastausta. *Astmatupakka* –hakusanalla tuli hakutulos: ei vastaa yhtään sivua. Koska kysymyksessä ei mainittu tupakka –sanaa, oli haettava pelkillä *astma* ja *apteekki* –hakusanoilla.

Toinen laajempi iGS-kategoria mihin ei oikein löytynyt Googlella vastauksia oli kielet, kielitiede, sanat. Vain joka viidenteen (20,8 prosenttia) löytyi vastaus. Osassa tämän kategorian kysymyksissä tiedusteltiin sanojen tai sanontojen alkuperää. Luokittelun kannalta tällaiset kysymykset osoittautuivat ongelmallisiksi. Pelkän kysymyksen perusteella oli mahdotonta luokitella kysymys joko fakta- tai aihekysymykseksi: luokittelussa oli kysymyksen lisäksi perehdyttävä myös (alkuperäiseen) vastaukseen. Mikäli sanonnan alkuperä oli selkeästi osoitettavissa, oli kyse faktakysymyksestä. Muussa tapauksessa (kun esim. sanonnan alkuperästä oli monia tulkintoja), kysymys luokiteltiin tiettyä aihetta koskevaksi kysymykseksi.

Esimerkki 13

> *Sanotaan, että ”On lottovoitto syntyä Suomeen”. Mistä tämä sanonta on saanut alkunsa?*

Kysymyksestä voidaan poimia hakusana *sanonta* sekä fraasi *On lottovoitto syntyä Suomeen*. Erisnimet ja tällaiset sanonnat, joita ei ole järkevää pilkkoa osiin, käsitellään fraaseina, ja lasketaan yhdeksi hakukäsitteeksi. Tällaisissa tapauksissa on sallittua ylittää ”simulaation” kolmen hakusanan rajaus.

Google-hakujen tuloksina oli ainoastaan joukko erilaisia blogeja ja keskustelupalstoja, missä oli mainittu kyseinen sanonta eri yhteyksissä.

4.5 Tiettyä teosta koskevat kysymykset

Suurin osa tiettyä teosta koskevista kysymyksistä liittyi musiikkiin (25 kpl eli 40,3 prosenttia). Kysyjät halusivat tietää laulun nimiä ja/tai esittäjiä, laulun sanoituksia, tai sitten vain halusivat tietää että mistä laulu löytyisi. Muutama kysymys liittyi lisäksi laulun sisältöön.

Useimmiten kysyjä tiesi joko laulun esittäjän tai kappaleen nimen (tai molemmat). Toisinaan kysyjä muisti ainoastaan pienen pätkän laulun sanoista, jonka perusteella pitäisi tunnistaa kyseinen kappale.

Tiettyä teosta koskevien kysymysten osalta toiseksi suurin iGS-kategoria oli elokuva, video, teatteri, tanssi (14 kpl eli 22,6 prosenttia). Suurin osa näistä kysymyksistä liittyi elokuviin (10 kpl). Kysyjät halusivat esim. tietää että milloin elokuva tulee elokuvateattereihin, tai mistä elokuva löytyisi. Muutama kysymys liittyi elokuvien sisältöön.

Seuraavaksi eniten teoksiin liittyvistä kysymyksistä tuli TV, radio, lehdet –kategoriasta (9 kpl eli 14,5 prosenttia). Suurin osa näistä kysymyksistä koski TV-ohjelmia. TV-ohjelmista haluttiin tietää mm. lähetetäänkö tv-ohjelma joskus uusintana, tai milloin ohjelma on tullut tv:stä.

Kirjoihin ja kirjallisuuteen liittyviä kysymyksiä esitettiin myös 9 kpl (14,5 prosenttia). Kysymykset liittyivät lähes poikkeuksetta kirjoihin (7 kpl). Usein kysymys liittyi teoksen sisältöön (esim. miten

hahmo on suomennettu), tai sitten kysyttiin kirjaa jostain tietystä aiheesta (esim. suljetun huoneen mysteerit) Loput tiettyä teosta koskevista kysymyksistä koskivat artikkeleita, musiikkivideoita tai runoja. Joukossa oli myös yksi julistetta koskeva kysymys. Näissä kysymyksissä kysyttiin useimmiten tekijää, tai sitä mistä teos löytyisi.

Kysymykset joihin löytyi vastaus Googlella

Koska tiettyä teosta koskevien kysymysten määrä oli näinkin vähäinen (yhteensä 62 kpl), ei voida tehdä kovinkaan pitkälle meneviä johtopäätöksiä Google-hakutulosten suhteen. Teosaiheisiin kysymyksiin oli vaikeinta löytää suomenkielistä relevanttia vastausta: Googlella löytyi vastaus ainoastaan 18 kysymykseen (29 prosenttia). Elokuva-aiheisiin kysymyksiin löytyi suhteellisesti eniten vastauksia (puolet), ja musiikkikappaleisiin liittyviin kysymyksiin löytyi myös melko hyvin vastauksia (10 kpl 25:stä).

Esimerkki 14

> *Kuka esitti suomeksi Jugoslavian euroviisun v. 1983 Dzuli / Julie?*

Hakulausekkeella *Dzuli Julie* löytyi Ylen äänitearkiston linkki, mistä löytyi oikea vastaus.

Esimerkki 15

> *Miten menee laulun Pikku Lauri sanat? Laulu alkaa, että Maailma on niin lavea pikku pikku Lauri…*

Hakulauseella *Pikku Lauri sanat* löytyi pdf-dokumentti, missä oli kyseisen laulun sanat sointumerkein.

Esimerkki 16

> *Milloin uusi "Jäniksen vuosi (Vatasen jänis)" ja "Edit Piaf (Pariisin varpunen)" tulevat Suomessa elokuvateattereihin?*

Vastaus löytyi molempiin kysymyksiin. Hakusanoilla *Jäniksen vuosi* löytyi Espoon Cinén nettisivu missä mainostettiin Espoon elokuvajuhlia. Nettisivuilla kerrottiin, että ”Arto Paasilinnan romaanista

Jäniksen vuosi tehty ranskalainen filmatisointi Le lièvre de Vatanen esitetään Espoon elokuvajuhlilla eli Espoo Cinéssä elokuun lopulla.” Samalla sivulla oli myös tarkempi päivämäärä esitykselle (21.-26. elokuuta).

Hakusanoilla *Edit Piaf* löytyi Finnkinon nettisivu missä kerrottiin Pariisin varpunen – Edit Piaf –elokuvasta, ja mainittiin myös elokuvan ensi-illan ajankohta (30.3.2007).

Kysymykset joihin ei löytynyt Googlella vastausta

Googlella ei kuitenkaan löytynyt vastausta suurimpaan osaan (42 kpl) tiettyä teosta koskevaan kysymykseen. TV-ohjelmia koskevat kysymykset olivat Google-hakujen suhteen hankalimpia: 80 prosenttia näistä kysymyksistä oli sellaisia joihin ei löytynyt Googlella vastausta.

Esimerkki 17

> *Terve! Tv-ohjelmista pari kiperää joihin en mistään löytänyt itse vastausta. milloin on tullut tv:stä (kanavasta ei tietoa) ohjelma "La Pigallen stripparit"? Vuosi ja päivämäärä olisi kiva tietää. Ja minkä nimisessä jaksossa Paavo Pesusienellä oli kotitalossaan isot teknojuhlat?*

La Pigallen stripparit –hakulausekkeella ei löytynyt kuin vaillinainen hakutulos (2 dokumenttia), eikä iGS-vastaajakaan löytänyt kyseistä ohjelmaa, joten ilmeisesti kysyjä muisti nimen väärin. *Paavo Pesusieni teknojuhlat* –hakulausekkeella ei tullut lainkaan hakutulosta, eikä *Paavo Pesusieni* –haullakaan löytynyt kyseistä teknojuhlajaksoa (kaikkea muuta kyllä, kuten Paavo Pesusieni –pelejä).

Esimerkki 18

> *Haluaisin tietää mistä löytyisi TERTTU valssi olen kuullut että on levytetty kiitos*

Koska kysymyksestä ei voinut poimia muita kuin *Terttu* ja *valssi* –hakusanat, ja molemmat hakusanat olivat välttämättömiä haun kannalta, en voinut tehdä kuin yhden haun. Hakusanoilla *Terttu valssi* löytyi enimmäkseen sekalaisia sivuja, joissa Terttu ja valssi olivat toki mainittu, mutta eivät liittyneet mitenkään toisiinsa (esim. Someron tapahtumia, sukuseuran jäsenkirje, Dance World Cup:in tulokset, kirjojen nettikauppa jne.). Sama ongelma oli myös hakutuloksen muutamissa äänitteisiin liittyvissä sivuissa.

5. YHTEENVETO JA JOHTOPÄÄTÖKSET

Tämän tutkielman päätavoitteena oli selvittää, missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS-verkkotietopalvelun kysymyksiin, ja minkätyyppisiä kysymyksiä asiakkaat lähettävät iGS-palveluun. Lisäksi tutkimuksessa selvitettiin, onko iGS:n kysymysten tyypillä yhteyttä siihen että missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin.

Tutkimusmetodina käytettiin tätä tutkimusta varten kehitettyä hakukoneiden keskivertokäyttäjien ”simulaatiota”. Hakukoneiden käyttäjistä saatua tutkimustietoa sovellettiin eräänlaisten tutkimuksen ennakkoehtojen ja rajausten luomisessa. Käyttäjätutkimusten tulosten perusteella muodostettiin hakukoneiden keskivertokäyttäjiä ”simuloivia” sääntöjä, joiden tarkoituksena oli vähentää hakijan yksilöllisten ominaisuuksien vaikutusta hakutuloksiin.

”Simulaation” tärkeimmät säännöt olivat: 1) kyselyjen hakutermien määrä rajataan maksimissaan kolmeen hakutermiin, 2) edistyneempiä hakutekniikoita ei käytetä, 3) kyselyissä käytetyt käsitteet poimitaan alkuperäisistä iGS-kysymyksistä, 4) käsitteitä edustavat hakutermit muutetaan nominatiivi- eli perusmuotoon, 5) kyselyt muodostetaan systemaattisella menetelmällä, missä tärkeimpiä pääkäsitteitä edustavia hakutermejä yhdistetään muihin hakutermeihin, 6) hakuja tehdään korkeintaan viisi kappaletta yksittäistä alakysymystä kohden, 7) kyselyt tehdään suomen kielellä ja 8) hakutuloksista käydään läpi ainoastaan ensimmäisen tulossivun dokumentit (10 kpl).

Suuri osa edellä mainituista säännöistä perustui hakukoneiden käyttäjätutkimusten tuloksiin. Valitettavasti täydellistä hakukoneiden keskivertokäyttäjän ”simulaatiota” ei pystytty toteuttamaan pelkästään tutkimustulosten perusteella. Sääntöjä muodostaessa jouduttiin tekemään jonkin verran oletuksia hakukoneiden keskivertokäyttäjän hakutavoista. Lisäksi monimutkaiset ja monisanaiset hakupyynnöt osoittautuivat lopulta varsin haastaviksi sääntöjen muodostamisen ja tutkimuksen objektiivisuusvaatimuksen kannalta. Hakukäsitteiden valinnasta ja hakulausekkeiden muodostamisesta muodostui joissain tapauksissa varsin subjektiivinen prosessi. Toisaalta nämä valintoihin liittyvät subjektiivisuusongelmat eivät olleet ratkaisevia tutkimusasetelman kannalta. ”Simulaation” ideana kun oli osoittaa, että siinä esitettyjen sääntöjen puitteissa on kenen tahansa hakukoneiden keskivertokäyttäjän mahdollista päätyä *vähintään* samaan määrälliseen tulokseen.

Tutkimusmenetelmällä saatiin siis aikaiseksi eräänlainen vähimmäistulos, ja mahdollisten puutteiden vaikutus hakutuloksiin oli yhdensuuntainen: vastauksia olisi voinut löytyä enemmänkin.

”Simulaation” avulla Google-hakukoneella löytyi relevantti suomenkielinen vastaus 188 kysymykseen. Tutkimuksen otoksen 509 alakysymyksestä tämä oli 36,9 prosenttia, eli runsas kolmasosa. Kun ottaa huomioon kuinka paljon ”simulaation” sääntöjen avulla hakujen muodostamista rajoitettiin, tämä tuntuu varsin suurelta luvulta. Tämän tutkimuksen mukaan Helsingin kaupunginkirjaston iGS-verkkotietopalveluun lähetetyistä kysymyksistä runsas kolmasosa on siis sellaisia, joihin periaatteessa kuka tahansa hakukoneiden keskivertokäyttäjä voisi löytää melko helposti vastauksen. Itse toimin nykyisin iGS-vastaajana, ja yllätyin melko lailla tästä tuloksesta. Päällimmäiset ajatukseni olivat: Miksi palveluun lähetetään näin paljon ”helppoja” kysymyksiä? Onko joidenkin hakijoiden hakutaidoissa pahoja puutteita, vai ovatko hakijat vain laiskoja, eivätkä itse viitsi vaivautua ”googlettamaan”? Näihin kysymyksiin ei tässä tutkimuksessa etsitty vastausta.

Toisaalta suurin osa (63,1 prosenttia) kysymyksistä oli sellaisia, joihin ei löytynyt vastausta Googlella. Palveluun lähetetään kuitenkin runsaasti sellaisia kysymyksiä, joihin ei ihan helposti löydy ”googlettamalla” vastausta. Tutkimuksessa ei erikseen selvitetty miksi vastausta ei löytynyt, mutta alkuperäisten vastausten lähteistä pystyi jonkin verran päättelemään mahdollisia syitä. Niitä olivat mm. seuraavat: hakupyynnössä ei ollut riittävästi sopivia hakukäsitteitä kunnollisten hakujen suorittamiseen, tietoa ei löytynyt suomenkielisistä lähteistä (olisi pitänyt hakea englanninkielisillä hakusanoilla), ja tietoa ei löytynyt ylipäänsä netistä (vaan vastaus löytyi joko kirjallisuudesta tai asiantuntijalta).

Tutkimuksessa selvitettiin myös että minkätyyppisiä kysymyksiä asiakkaat lähettivät iGS-verkkotietopalveluun. Otoksen kysymykset jaettiin kolmeen kysymyskategoriaan. Eniten palveluun lähetettiin tiettyä aihetta koskevia kysymyksiä (44,2 prosenttia), mutta lähes yhtä paljon lähetettiin faktakysymyksiä (43,6 prosenttia). Selvästi vähiten oli tiettyä teosta koskevia kysymyksiä (12,2 prosenttia). Vaikka Karinen (2008) analysoi tutkimuksessaan iGS-palvelun kysymystyyppejä, eivät hänen käyttämänsä luokittelukategoriat olleet vertailukelpoisia tämän tutkimuksen kanssa. Sitä vastoin Gräsbeckin (2008) Kysy kirjastonhoitajalta –palvelun kysymystyyppien luokittelu oli tärkeimpien luokkien osalta identtinen tämän tutkimuksen kanssa. Gräsbeckin tutkimuksessa Kysy kirjastonhoitajalta –palveluun lähetettiin eniten tiettyä teosta koskevia kysymyksiä (33 prosenttia).

Tiettyä aihetta koskevia kysymyksiä oli toiseksi eniten (25 prosenttia), ja lähes yhtä paljon oli faktakysymyksiä (23 prosenttia). Suurimmat eroavaisuudet tulivat tiettyä teosta koskevien kysymysten osalta. Tämä vahvistaa omalta osaltaan aikaisempaa käsitystä siitä, että Kysy kirjastonhoitajalta –palvelusta kysytään enemmän kirjastoon ja sen kokoelmiin liittyviä kysymyksiä, kun taas iGS-palvelussa kirjastomaisuus on jätetty taka-alalle.

Myös Høivikin (2005) tutkimuksessa käytettiin vastaavia kategorioita, joten tulokset ovat siltä osin vertailukelpoisia. Norjalaiseen Kysy kirjastosta –palveluun lähetettiin eniten aihekysymyksiä (noin puolet), toiseksi eniten dokumenttikysymyksiä (noin kolmasosa), ja vähiten faktakysymyksiä (noin kuudesosa). Toisaalta Høivikin otos oli varsin pieni (100 kysymystä), joten palvelujen kysymystyyppien eroavaisuuksista ei kannata tehdä kovin pitkälle meneviä johtopäätöksiä.

Lisäksi tutkimuksessa yritettiin selvittää, onko iGS:n kysymysten tyypillä yhteyttä siihen että missä määrin Google-hakukoneella löytyy relevantteja suomenkielisiä vastauksia iGS:n kysymyksiin. Tilastollisesti merkitsevää yhteyttä muuttujien välillä ei kuitenkaan todettu.

Hakukoneiden keskivertokäyttäjien ”simulaatiosta” saatuja tuloksia on melko vaikeaa yleistää, koska tällaista tutkimusmetodia ei ole koskaan aikaisemmin käytetty. Tutkimusmetodilla saatiin kuitenkin suuntaa-antava ”vähimmäistulos”, ja lisäksi saatiin tietoa palvelun kysymystyypeistä. Tutkimustulosten yleistettävyys vaatisi laajempaa tutkimusaineistoa ja ”simulaation” sääntöjen kehittämistä edelleen. ”Simulaatio” toteutettiin vuonna 2007, ja säännöt perustuivat sitä ennen tehtyihin hakukoneiden käyttäjätutkimuksiin. Tuoreempien tutkimusten avulla olisi mahdollista tehdä tarkempia hakukoneiden keskivertokäyttäjiä kuvaavia sääntöjä.

Uudenlaisen tutkimusmetodin luominen melko lailla tyhjästä oli varsin vaativa ja aikaa vievä tehtävä, mutta myös mielenkiintoinen ja opettavainen prosessi. Sääntöjä kehittäessä jouduin toden teolla pohtimaan tiedonhaun prosessia ja siihen liittyviä asioita, sekä ratkaisemaan eteen tulevia ongelmia parhaan kykyni mukaan. ”Simulaation” sääntöjen kehittämisessä jouduin tekemään jonkin verran kompromisseja (muutamat oletukset hakukoneiden keskivertokäyttäjän hakutavoista sekä subjektiivisten valintojen mahdollisuus), mutta sain aikaiseksi melko loogiselta tuntuvan kokonaisuuden. ”Simulaation” sääntöjen noudattaminen oli suhteellisen helppoa. Varsinaisten Google-hakujen tekeminen ”simulaation” sääntöjä noudattamalla oli oikeastaan erittäin mielenkiintoista, toisinaan jopa hauskaa (vaikkakin aikaa vievää).

Kehittämääni tutkimusmenetelmää, hakukoneiden keskivertokäyttäjien ”simulaatiota”, voidaan kehittää edelleen, ja soveltaa erilaisissa tiedonhakuun liittyvissä tutkimuksissa. Olisi esimerkiksi mielenkiintoista tutkia, minkälaisia lähteitä iGS-vastaajat käyttävät vastauksissaan, ja olisiko käytetyillä lähteillä yhteyttä Google-hakujen avulla saatuihin hakutuloksiin. Tutkimusmenetelmää voitaisiin käyttää myös muiden verkkotietopalvelujen tutkimisessa.

LÄHTEET

Alaterä, A. & Halttunen, K. 2002. Tiedonhaun perusteet – osa lukutaitoa. Helsinki: BTJ Kirjastopalvelu.

Googlen verkkosivut. <http://www.google.com> (Käytetty 15.8.2008)

Granlund, N. & Lönn, R. 2007. Selvitys Helsingin kaupunginkirjaston verkkotietopalvelun uudelleenorganisoinnista: iGS ja Kysy kirjastonhoitajalta. Julkaisematon.

Gräsbeck, T. 2008. Kysy kirjastonhoitajalta –verkkotietopalvelun vastausarkiston kysymysten analyysi. Tampereen yliopisto. Informaatiotutkimuksen laitos. Pro gradu –tutkielma.

Hsieh-Yee, I. (1993). Effect of search experience and subject knowledge on online search behavior: Measuring the search tactics of novice and experienced searchers. Journal of the American Society for Information Science 44(3), 161–174. Saatavilla sähköisessä muodossa http://www.asis.org/Publications/JASIS/Best_Jasist/1994Hsieh-Yee.pdf Haettu 4.10.2006

Høivik, T. 2005. A Poem lovely as a tree? Virtual reference questions in Norwegian public libraries. Teoksessa Johannsen, C.G. & Kajberg, L. (toim.) New frontiers in public library research. Lanham, Md.: Scarecrow Press, 43–59.

Hälinen, J. 2004. Kirjastojen verkkotietopalvelujen arviointi palvelun tuottajan ja käyttäjän näkökulmasta. Koulutus- ja kehittämiskeskus Palmenia. Helsingin yliopisto. Proseminaarityö.

Hölscher, C. 1998. How Internet experts search for information on the Web. Teoksessa Maurer, H. & Olsen, R. G. (toim.) Proceedings of WebNet98 - World conference of the WWW, Internet & Intranet. Association for the Advancement of Computing in Education. Saatavilla sähköisessä muodossa http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/17/5a/11.pdf Haettu 15.6.2006

Hölscher, C. & Strube, G. 2000. Web search behavior of Internet experts and newbies. Computer Networks, 33(1), 337–346.

iGS Tietohuoltoaseman verkkosivut. <http://igs.kirjastot.fi/> (Käytetty 20.1.2010)

iGS Tietohuoltoaseman projektiraportti 2001–2002. Julkaisematon.

iGS Tietohuoltoaseman toimintakertomus 2007. Julkaisematon.

Iivonen, M. 1995. Hakulausekkeiden muotoilun yhdenmukaisuus onlineviitehaussa. Tampere: Tampereen yliopisto (Acta Universitatis Tamperensis; A443).

Ihamäki, S. 1999. Etäistä? Kysy kirjastonhoitajalta –etätietopalvelun tarkastelua. <http://pandora.lib.hel.fi/julkaisut/etaista/> (Käytetty 5.11.2006)

Ihamäki, S. & Juntumaa, J. 2002. Kysymällä paras. Kirjastolehti 95(4), 16–17.

Jansen, B.J. (2000). The effect of query complexity on Web searching results. Information Research, 6(1). Saatavilla sähköisessä muodossa http://informationr.net/ir/6-1/paper87.html Haettu 8.6.2006

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users and real needs: A study and analysis of users' queries on the Web. Information Processing and Management, 36(2), 207-227. Saatavilla sähköisessä muodossa http://www.searchlores.org/library/ipm98.pdf Haettu 15.8.2006

Juntumaa, J. 1998. Mitä etätietopalvelusta kysytään ja mitä vastataan. <http://www.lib.hel.fi/julkaisut/vtp.htm> (Käytetty 5.11.2006)

Juntumaa, J. 2003. Chat-neuvonnan kokeilu Helsingin kaupunginkirjastossa. <http://pandora.lib.hel.fi/julkaisut/chat-raportti.pdf> (Käytetty 21.1.2010)

Juntumaa, J. 2004. Helsingin kaupunginkirjaston verkkotietopalvelu: sisällönanalyysiä. <http://www.lasipalatsi.fi/suomenkirjasto/chat/verkkotietopalvelun_sisalto.ppt> (Käytetty 13.10.2006)

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

Järvelin, K. & Sormunen, E. 1999. Dokumentit kateissa? Tiedon tallennus ja haku avuksi. Teoksessa Mäkinen, I. (toim.). Tiedon tie: Johdatus informaatiotutkimukseen. Helsinki: BTJ Kirjastopalvelu, 117.

Karinen, V. 2008. Saa kysyä! iGS Tietohuoltoaseman kysymysten analyysi. Tampereen yliopisto. Informaatiotutkimuksen laitos. Pro gradu –tutkielma.

Lazuly, Pierre. 2007. Maailma Googlen mukaan. Le Monde Diplomatique IV. 2007. Helsinki: Voima, 105–112.

Lucas, W. & Topi, H. (2002). Form and function: The impact of query terms and operator usage on web search results. Journal of the American Society for Information Science and Technology. 53(2), 95–108. Saatavilla sähköisessä muodossa http://www.ece.uc.edu/~annexste/Courses/cs690/formandfunction.pdf Haettu 4.12.2006

Numminen, P. 2008. Kysy kirjastonhoitajalta neuvontapalvelun kysymystyypit vuonna 1999 ja 2006. Tampereen yliopisto. Informaatiotutkimuksen laitos. Pro gradu –tutkielma.

Nurmi, T., Rekiaro, I. & Rekiaro, P. 1993. Suomea suomeksi : suomen kielen sanakirja. Jyväskylä: Gummerus.

Pomerantz, J. 2005. A linguistic analysis of question taxonomies. Journal of the American Society for Information Science and Technology 56(7), 715–728. Saatavilla sähköisessä muodossa http://www.ils.unc.edu/~jpom/pubs/Preprint-JASIST-2005a.pdf Haettu 5.10.2006

Sanastokeskus TSK ry:n verkkosivut, TEPA-termipankki. <http://www.tsk.fi> Viitattu 21.1.2010

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1999). Analysis of a very large Web search engine query log. SIGIR Forum 33(1), 6 – 12. Saatavilla sähköisessä muodossa http://www.cs.ucsb.edu/~almeroth/classes/tech-soc/2005-Winter/papers/analysis.pdf Haettu 2.11.2006

Spink, A. & Jansen, B. (2004). Web search: public searching of the web. Dordrecht: Kluwer Academic Publishers.

Tietotekniikka- ja verkkopalvelustrategia 2003–2006. Julkaisematon.