Raija Lehtokangas

Sanomalehtien yhdenmukaisuus sananvalinnassa samoista uutisaiheista kirjoitettaessa

Empiirinen tutkimus 180 uutisjutusta

Raija Lehtokangas, Sanomalehtien yhdenmukaisuus sananvalinnassa samoista uutisaiheista kirjoitettaessa - empiirinen tutkimus 180 uutisjutusta [Word consistency in news documents on the same topics - an empirical study on 180 newspaper documents]. Informaatiotutkimus 19 (3): 82-89, 2000.

This article investigates how consistent different newspapers are in their choice of words when writing about the same topics. News articles on same topics were taken from three Finnish newspapers and compared in regard to their central concepts and words representing the concepts in the news texts. Consistency figures were calculated for each set of three articles (the total number of them was 60). Inconsistency in words and concepts was found between news articles from different newspapers. The mean value of consistency calculated on the basis of words was 64.8 %; the concept consistency was considerably higher. The news articles represented two categories of length and three categories of topic (event, process and opinion) and consistency was studied separately for each category of length and topic. Statistically significant differences in consistency were found according to the categories of length but not to the categories of topic.

Address: Raija Lehtokangas, Sepänkatu 10 A 12, FIN-33230 Tampere, Finland

Johdanto

Artikkeli pohjautuu Tampereen yliopiston informaatiotutkimuksen laitokselle tekemääni pro gradu tutkielmaan *Sananvalinnan yhdenmukaisuus sanomalehtiuutisissa* (Lehtokangas 1999). Tutkielmassa tarkasteltiin sitä, kuinka yhdenmukaisia sanomalehdet ovat sananvalinnaltaan samoista uutisaiheista kirjoittaessaan, ja sen tarkoituksena oli omalta osaltaan lisätä tietämystä luonnollisella kielellä kirjoitettujen tekstien sananvalinnan yhdenmukaisuudesta. Yhdenmukaisuudella (engl. consistency) tarkoitetaan sitä, kuinka yhtenevällä tavalla toimitaan suoritettaessa samaa tehtävää eri tilanteissa - joko niin, että tarkastellaan useaa toimijaa yhdessä tilanteessa (toimijoiden välinen yhdenmukaisuus) tai niin, että tarkastellaan yhtä toimijaa useassa tilanteessa (toimijan sisäinen yhdenmukaisuus). Yhdenmukaisuudella mitataan suoritusten samanlaisuutta ottamatta kantaa niiden laadukkuuteen. (Iivonen 1995, 11) Tutkimuksessani yhdenmukaisuustarkastelua sovellettiin sanomalehtijuttujen teksteihin - siihen, miten samanlaisia ne ovat sanastoltaan. Minkä tahansa tekstin voidaan ajatella sisältävän informaatiota käsiterakenteena, joka voidaan kuvata käsitteinä ja niiden välisinä suhteina. Käsitteet puolestaan ilmaistaan tekstin tasolla esimerkiksi luonnollisen kielen sanojen

avulla. (Järvelin 1995, 68-71) Myös omassa tutkimuksessani yhdenmukaisuutta tarkasteltiin sekä ilmaisujen että käsitteiden osalta.

Informaatiotutkimuksen alalla on aiemmin tutkittu indeksoinnissa ja tiedonhaussa käytettyjen termien ja käsitteiden yhdenmukaisuutta sekä hakutulosten yhdenmukaisuutta. Pääpaino on ollut eri toimijoiden välisessä yhdenmukaisuudessa, ja toimijoiden sisäisen yhdenmukaisuuden tutkiminen on ollut vähäisempää. Niin ikään tutkimuksissa on yleensä keskitytty ilmaisutason yhdenmukaisuuteen ja käsitetason tarkastelu on jäänyt vähemmälle huomiolle. (Iivonen 1993, 64-65) Tutkimuksessani yhdenmukaisuustarkastelua sovellettiin tekstin sisältämiin ilmaisuihin ja niiden taustalla oleviin käsitteisiin. Tiedossani ei ole, että yhdenmukaisuustutkimusta olisi aikaisemmin käytetty tähän tarkoitukseen, joten kyseessä lienee menetelmän soveltaminen uudella alueella.

Tutkimuksen aihepiirinä on tiedonhaku tekstitietokannoista. Tekstitietokantojen määrä on kasvanut nopeasti (Gale directory of databases 1997). Tekstikannoista puuttuu usein indeksointi tai muu sisällönkuvailu. Tällöin dokumenttien sanaston yhdenmukaisuudella on suora vaikutus tiedonhakuun: mitä enemmän sanasto vaihtelee dokumentista toiseen, sitä enemmän tiedonhakijan on käytettävä haussaan vaihtoehtoisia ilmaisuja hyvään hakutulokseen päästäkseen. Ihmisten välisessä viestinnässä käytettävälle kielelle on tyypillistä, että sama asia voidaan ilmaista usealla eri tavalla. Kielellistä vaihtelua voi esiintyä kaikilla kielen tasoilla - tämän tutkimuksen kannalta kiinnostavinta on sanastoon liittyvä vaihtelu. Tutkimuksen kohde, sanomalehden uutisjutut, on dokumenttilaji, jonka rakenne on vakiintunut tietynlaiseksi ja kieli on parhaimmillaan selkeää, helposti ymmärrettävää käyttökieltä (Miettinen 1984, 74-76; Okkonen 1986, 209-233).

Tutkimusongelma

Tutkimuksen tarkoituksena oli selvittää, kuinka yhdenmukaisia sanomalehdet ovat sananvalinnaltaan samoista aiheista kirjoittaessaan. Tätä varten otettiin kolmesta eri lehdestä samoihin uutisaiheisiin liittyviä juttuja ja selvitettiin, kuinka yhdenmukaista sananvalinta on samaan uutisaiheeseen liittyvissä kolmen eri lehden jutuissa. Koska sananvalinnan vaihtelu lehtien välillä saattaa olla erilaista erilaisissa uutisjutuissa, aineistoa ei tarkasteltu vain yhtenä kokonaisuutena vaan myös osiin jaettuna. Jakoperusteita oli kaksi: uutisjutun laajuus ja uutisjutun tyyppi, joka määräytyi uutisen[1] sisällön mukaan. Laajuudeltaan uutiset kuuluivat kahteen luokkaan: sähkeuutisissa oli korkeintaan 100 sanaa, laajoissa uutisissa vähintään 200 sanaa. Uutisia oli kolmea eri tyyppiä: tapahtuma-, prosessi- ja kannanottouutisia; nimitän niitä uutistyypeiksi. Kannanottouutisen pääosan muodosti henkilön, organisaation tms. kannanotto johonkin tai joihinkin aiheisiin, kyseessä oli esimerkiksi puheen referointi. Tapahtumauutiseen sisältyi aina keskeisesti hetki, jonka jälkeen asiat olivat toisin kuin ennen, tämä oli esimerkiksi hetki, jolloin sattui onnettomuus tai muu yllättävä tapahtuma, saavutettiin tulos, tehtiin päätös, päästiin sopimukseen tms. Tapahtuma itsessään oli lyhytkestoinen, tai se esitettiin sellaisena esimerkiksi lopputulosta korostamalla (esim. uutisessa vallankaappauksesta huomio oli lopputuloksessa: valta on siirtynyt uusille johtajille, edeltäviä tapahtumia ei kuvata). Juttuja, jotka eivät täyttäneet tapahtuma- tai kannanottouutisen kriteerejä, pidettiin prosessiuutisina. Niissä kuvattiin toimintaa pitkän tai rajoittamattoman ajan kuluessa tai aikaperspektiivi puuttui kokonaan, jolloin kyseessä oli vallitsevan tilanteen kuvailu (esim. Saksan työllisyystilanne tammikuussa 1996). Monesti tapahtuma- ja prosessiuutisen välillä ei ollut suurta eroa, vaan jo varsin vähäisillä muutoksilla uutinen olisi voitu siirtää toiseen ryhmään.

Käyttämällä sekä laajuutta että tyyppiä jakoperusteena saatiin aineisto jaetuksi kuuteen ryhmään, joissa kussakin yhdenmukaisuutta voitiin tarkastella erikseen. Luonnollisesti yhdenmukaisuutta voitiin tarkastella myös ryhmissä, jotka oli saatu käyttämällä vain jompaakumpaa jakoperustetta. Yhdenmukaisuutta tarkasteltiin kahdella tasolla. Ilmaisujen yhdenmukaisuusarvo kuvasi tutkimuksen aiheena olevaa sananvalinnan yhdenmukaisuutta. Käsitetason tarkastelu oli mukana, jotta saataisiin selville, johtuiko epäyhdenmukaisuus sananvalinnassa siitä, että käsitteille oli eri lehdissä valittu eri ilmaisut vai siitä, että käsitteet itsessään olivat erilaisia eri lehdissä.

Aineisto ja menetelmät

Tutkimuksen aineistona oli 180 uutisjuttua, jotka olivat peräisin kolmesta sanomalehdestä: Aamulehdestä (AL), Helsingin Sanomista (HS) ja Turun

[1]Sanoja uutinen ja uutisjuttu käytetään synonyymeinä tässä tutkimuksessa.

Sanomista (TS). Jokaisesta lehdestä oli 60 juttua, ja ne jakaantuivat tasan sähkeuutisiin ja laajoihin uutisiin. Lisäksi jutut edustivat kolmea tyyppiä: tapahtuma-, prosessi- ja kannanottouutisia (kutakin tyyppiä 20 kpl). Suurin osa jutuista oli vuoden 1996 neljältä ensimmäiseltä kuukaudelta. Aineistoa hankittaessa oli ensimmäisenä tehtävänä löytää uutisaiheita, joista löytyisi kriteerit täyttävä uutisjuttu kaikista kolmesta lehdestä. Kun sopiva juttukolmikko oli löytynyt, käsiteltiin jokainen jutuista erikseen: etsittiin keskeiset käsitteet, keskeisiä käsitteitä vastaavat ilmaisut, perusmuotoistettiin ilmaisut ja lopuksi laskettiin kolmen jutun välinen yhdenmukaisuusarvo sekä ilmaisujen että käsitteiden osalta. Tutkimuksen perusyksikkönä olivat siis juttukolmikot, joille laskettiin yhdenmukaisuusarvoja.

Keskeisiä käsitteitä etsittäessä ei ollut käytössä ennalta päätettyä käsitteiden joukkoa, vaan käsitteet määräytyivät viime kädessä kunkin jutun sisällön perusteella. Luonnollisesti mukana oli usein uutisten keskeisiä sisältökategorioita (ks. Niemi 1997, 32-50), kuten toiminta, paikka, toimija, kohde, syy, seuraus jne. Myös ilmaisuille asetettiin hyvin vähän rajoituksia etukäteen, ainoastaan joitakin sanaluokkia tai niiden osia suljettiin pois sillä perusteella, että niitä ei käytetä tiedonhaussa. Ilmaisut koostuivat yhdestä sanasta (ainoana poikkeuksena olivat nimet, ne saivat sisältää useita sanoja).

Samanaiheisten juttujen välisen yhdenmukaisuuden laskemiseksi käytettiin epäsymmetristä kaavaa (ks. esim. Iivonen 1993, 69-70). Saadut yhdenmukaisuusarvot ovat prosenttilukuja, jotka kuvaavat sitä, kuinka suuressa määrin vertailluissa jutuissa on käytetty samoja ilmaisuja ja käsitteitä. Arvo 100 tarkoittaa sitä, että kaikissa juttukolmikon jutuissa on täysin samat käsitteet tai ilmaisut, arvo 0 sitä, että kaikissa jutuissa on täysin eri käsitteet tai ilmaisut. Yhdenmukaisuusarvoja tarkasteltiin lähinnä keskiarvoina sekä koko aineiston tasolla että pienemmissä ryhmissä. Yhdenmukaisuusarvojen riippuvuutta uutisen laajuudesta ja tyypistä tutkittiin chi^2-riippumattomuustestillä. Ilmaisu- ja käsitetason yhdenmukaisuusarvojen keskinäistä riippuvuutta tutkittiin Pearsonin korrelaatiokertoimella.

Tulokset

Ilmaisujen yhdenmukaisuus

Koko aineistossa ilmaisujen keskimääräinen yhdenmukaisuus oli 64.8 (ks. Taulukko 1.). Sähke- ja laajojen uutisten keskiarvoissa oli suuri ero: edellisten keskiarvo oli 82.8, jälkimmäisten 46.8. Tutkimuksessa tuli ilmi kaksi seikkaa, jotka osaltaan vaikuttavat tähän eroon. Yksi niistä oli jutuissa käytetyn tekstin alkuperä. Laajojen uutisjuttujen tekstit olivat sähkeuutisia huomattavasti useammin toimituksen omia, sähkeuutisten tekstit taas olivat usein peräisin uutistoimistoista, millä oli varsin voimakas ilmaisuja yhdenmukaistava vaikutus. Lisäksi laajoissa jutuissa oli keskimäärin huomattavasti enemmän käsitteitä kuin sähkeissä - näin ollen mahdollisuus siihen, että eri lehtien jutuissa olisi eri käsitteitä ja sitä myöten myös eri ilmaisuja, oli laajoilla jutuilla suurempi. Tarkasteltaessa yhdenmukaisuutta uutistyypeittäin havaitaan, että tapahtuma- ja prosessiuutiset olivat keskimäärin lähes yhtä yhdenmukaisia (keskiarvot 62.3 ja 62.1), kannanottouutisten yhdenmukaisuus oli jonkin verran suurempaa (keskiarvo 70.1). Kannanottouutiset poikkesivat myös ilmaisujen määrissä muuntyyppisistä

Taulukko 1. Ilmaisujen yhdenmukaisuusarvojen keskiarvot tyypiltään ja laajuudeltaan erilaisissa uutisissa

		UUTISEN TYYPPI			
		Tapahtuma	**Prosessi**	**Kannanotto**	**Yhteensä**
UUTISEN LAAJUUS	**Sähke**	77,8	83,7	87,0	82,8
	Laaja	46.7	40,4	53,2	46,8
	Yhteensä	62,3	62,1	70,1	64,8

uutisista: ilmaisuja oli niissä keskimäärin vähemmän ja myös ilmaisujen määrä käsitettä kohti oli niillä pienempi.

Edellä yhdenmukaisuutta tarkasteltiin keskiarvoina. Yhdenmukaisuusarvojen vaihtelu ryhmien sisällä oli kuitenkin suurta, esimerkiksi laajoissa kannanottouutisissa pienin yhdenmukaisuusarvo oli 27.6 ja suurin 86.4. Näin ollen arvojen vaihteluvälit eli maksimi- ja minimiarvojen erotukset ryhmien sisällä muodostuivat usein varsin suuriksi (vaihteluväleistä pienin oli 34.6, suurin 58.8). Yhdenmukaisuusarvojen keskihajonta oli koko aineistossa 23.4; sähkeuutisten keskihajonta oli 14.8, laajojen uutisten 15.0.

Käsitteiden yhdenmukaisuus

Lehdet olivat käsitteiltään paljon yhdenmukaisempia kuin ilmaisuiltaan. Käsitteiden keskimääräinen yhdenmukaisuus koko aineistossa oli 94.6 (ks. Taulukko 2.). Myös arvojen vaihtelu oli vähäisempää kuin ilmaisuilla: keskihajonta oli koko aineistossa 5.4. Monet käsitteistä olivat varsin laajoja ja olisivat olleet jaettavissa edelleen käsitteiksi - käytännön syistä käsitteiden määrä juttua kohti oli kuitenkin pidettävä kohtuuden rajoissa. Jos käsitteitä olisi ollut käytössä enemmän, olisi juttujen välille ehkä saatu enemmän eroja ja niiden välinen yhdenmukaisuus olisi ehkä ollut nykyistä hieman alhaisempi.

Riippuvuuksien tutkiminen

Yhdenmukaisuusarvojen riippuvuutta uutisen laajuudesta ja tyypistä tutkittiin chi^2-riippumattomuustestillä. Tulokset olivat hyvin samansuuntaiset ilmaisu- ja käsitetasolla, molemmissa löydettiin käytetyllä riskitasolla (0.01) tilastollista riippuvuutta uutisen laajuuden ja yhdenmukaisuusarvon väliltä, mutta sen sijaan ei uutisen tyypin ja yhdenmukaisuusarvon väliltä.

Jokaiseen kolmen samanaiheisen uutisjutun kokonaisuuteen liittyi kaksi arvoa: yhdenmukaisuus ilmaisutasolla ja yhdenmukaisuus käsitetasolla. Näiden parittaisten arvojen korrelaatioksi saatiin Pearsonin korrelaatiokertoimella mitattuna 0.6546 koko aineistossa. Käsite- ja ilmaisutason yhdenmukaisuuden välillä vallitsi siis melko voimakas positiivinen korrelaatio. Selitysasteeksi saatiin 42.8, eli 42.8 % ilmaisujen yhdenmukaisuuden vaihtelusta selittyi käsitteiden yhdenmukaisuuden vaihtelulla; loput johtui jostakin muusta.

Esimerkkejä epäyhdenmukaisuudesta

Tutkimuksessa esille tulleet epäyhdenmukaisuustapaukset olivat keskenään hyvin erilaisia. Osa

Taulukko 2. Käsitteiden yhdenmukaisuusarvojen keskiarvot tyypiltään ja laajuudeltaan erilaisissa uutisissa

		UUTISEN TYYPPI			
		Tapahtuma	**Prosessi**	**Kannanotto**	**Yhteensä**
UUTISEN LAAJUUS	**Sähke**	95,5	98,3	98,0	97,2
	Laaja	92,8	91,6	91,8	92,0
	Yhteensä	94,1	94,9	94,9	94,6

epäyhdenmukaisuudesta selittyi sillä, että eri lehdissä viitattiin selvästi eri kohteisiin todellisuudessa (esim. jostakin henkilöjoukosta mainittiin eri henkilöitä eri lehdissä) - on luonnollista, että käytetyt ilmaisutkin olivat tällöin erilaisia. Jatkossa tarkastellaan kuitenkin muuntyyppisiä tapauksia, sillä ne ovat tutkimuksen kannalta kiinnostavampia.

Toisiaan vastaavia ilmaisuja eli ilmaisuja, jotka tutkituissa uutisjutuissa välittivät saman asiasisällön (esim. viittasivat samaan todellisuuden kohteeseen), voidaan tarkastella esimerkiksi seuraavista näkökulmista:

1) millaisissa semanttisissa suhteissa ilmaisut ovat toisiinsa nähden,
2) kuinka samankaltaisia ilmaisut ovat ulkonaisesti.

Tarkastelen seuraavaksi kumpaakin ryhmää erikseen.

Semanttiset suhteet

Synonymia

Varsin luonnollinen suhde toisiaan vastaavien ilmaisujen välillä on synonymia. Synonymiaa on eriasteista. Jotkut ilmaisut ovat toistensa täysiä tai lähes täysiä synonyymeja, toiset taas osittaisia synonyymeja, jolloin ne ovat vain joissakin konteksteissa korvattavissa toisillaan. (Sivula 1989, 183-199; Häkkinen 1990, 76-77; Karlsson 1994, 203-204; Olkinuora 1992, 113-124)

Esimerkiksi

Onnettomuus sattui klo 18.52 Lahnajärvellä. (HS)
Turma sattui kello 18.50 Lahnajärvellä noin kilometri Helsinkiin päin. (AL)

Muun muassa seuraavat synonyymisten ilmaisujen tyypit olivat edustettuina aineistossani:

lyhenne - täydellinen ilmaisu:

tv-yhtiö (AL, HS)	-	televisioyhtiö (TS)
KOK (AL, TS)	-	Kansainvälinen olympiakomitea (HS)

sanaliitto - yhdyssana:

vallan käyttö (AL, TS)	-	vallankäyttö (AL, TS)
satunnainen erä (HS)	-	satunnaiserä (AL)

vierasperäinen ilmaisu - omaperäinen ilmaisu:

gallup (AL, TS)	-	mielipidetiedustelu (HS)
infrastruktuuri (AL, TS)	-	perusrakenne (HS)

nimien eri kirjoitustavat:

Jihye Ayash (AL)	-	Yahya Ayyash (TS)	-
Jahja Aijash (HS)			

Hierarkiasuhde

Toisiaan vastaavien ilmaisujen välillä voi olla myös hierarkkinen suhde (ks. hierarkkisten suhteiden tyypeistä esim. ISO 1986; Järvelin 1995, 151-152). Aineistossani oli esimerkkejä varsinkin hyponymiasta (ks. Karlsson 1994, 204-206). Esimerkiksi erääseen kerrostaloasuntoon viitattiin kuudella eri ilmaisulla, joista osa oli hyponymiasuhteessa keskenään. Ilmaisujen voidaan ajatella edustavan kolmea eri hierarkiatasoa seuraavasti:

Ylin taso:	asunto (AL, HS, TS)
Keskitaso:	yksiö (HS,TS); kerrostaloasunto (HS, TS); räjähdysasunto (HS)
Alin taso:	alakertahuoneisto (AL); räjähdysyksiö (HS)

Alakertahuoneistoa voidaan pitää kerrostaloasunnon hyponyymina, räjähdysyksiötä räjähdysasunnon hyponyymina jne.

Assosiaatiosuhde

Myös assosiaatiosuhde on mahdollinen toisiaan vastaavien ilmaisujen välillä (ks. assosiaatiosuhteiden tyypeistä esim. ISO 1986; Järvelin 1995, 150).

Esimerkiksi presidentti Clintonin puhetta käsitelleessä uutisessa laittomaan siirtolaisuuteen viitattiin toisessa lehdessä toiminnan, toisessa toimijoiden avulla:

Toiminta: (laiton) maahanmuutto (TS)
Toimija: (laiton) siirtolainen (AL)

Toisiaan vastaavien ilmaisujen samankaltaisuus

Toisiaan vastaavia ilmaisuja voidaan tarkastella myös siltä kannalta, kuinka samankaltaisia ne ovat puhtaasti ulkonaisesti. Epäyhdenmukaisuustapaukset erosivat toisistaan huomattavasti: toisen ääripään tapauksissa oli kysymys samoista ilmaisuista, joilla oli eri kirjoitustapoja, toisen pään tapauksissa oli täysin eri ilmaisut. Näiden väliin sijoittuivat tapaukset, joissa ilmaisut olivat osittain samat. Yleensä tällöin oli kysymys johdoksista tai yhdyssanoista, joissa oli yhteisiä osia. Tarkastelen seuraavaksi kumpaakin näistä ryhmistä erikseen.

Johdokset

Johdoksia tehdään liittämällä johtopäätteitä olemassaoleviin sanoihin. Tämä on hyvin produktiivinen sananmuodostuskeino suomessa (Lepäsmaa et al. 1996, 12, 14-15; Vesikansa 1978, 11-12, 17-18). Myös tutkituissa lehtijutuissa oli johdoksia käytetty runsaasti hyväksi, usein sen vuoksi että eri lehdissä oli erilainen lauserakenne, mikä edellytti eri sanaluokan sanojen käyttöä.

Esimerkiksi

Koneen on *tilannut* PT Surya Agung Kertas. (AL, TS)
Tilaaja on indonesialainen PT Surya Agung Kertas. (HS)

Jeltsin ja Lukashenko painottivat, että yhteisön jäsenet ovat edelleen *itsenäisiä* valtioita. (HS)

Kumpikin osapuoli säilyttää *itsenäisyytensä*, Jeltsin kertoi venäläisille ja valkovenäläisille. (AL, TS)

Yhdyssanat

Yhdyssanojen muodostaminen on suomessa hyvin vapaata (Häkkinen 1990, 144). Uutisjutuissa olikin paljon yhdyssanoja, osa niistä tilapäisiä, kyseistä tilannetta varten muodostettuja. Osa toisiaan vastanneista ilmaisuista oli yhdyssanoja, joilla oli yhteinen yhdysosa. Tavallisinta oli, että yhteinen osa oli sanojen jälkiosana. Yhteinen jälkiosa ilmaisi tällöin lajin tai luokan, johon kohde kuuluu; alkuosa oli valittu sen mukaan, mitä näkökulmaa haluttiin korostaa.

Esimerkiksi erääseen kerrostaloon viitattiin sanoilla

kerrostalo (AL, HS, TS)
vuokratalo (AL, HS, TS)
räjähdystalo (HS)
elementtikerrostalo (TS)

Varsin tavallista oli myös se, että sama sana esiintyi tutkituissa uutisjutuissa sekä itsenäisenä sanana että yhdyssanan osana. Puhuttiin esimerkiksi paitsi *sellun hinnasta* (AL,TS), myös *selluhinnasta* (TS), *tonnihinnasta* (TS) ja *maailmanmarkkinahinnasta* (HS). Yhdyssanojen alkuosat ilmaisivat tässä hyvin erilaisia asioita: kohdetta, jolle hinta kuuluu (sellu), yksikköä, jolle hinta lasketaan (tonni) ja paikkaa, jossa hinta maksetaan (maailmanmarkkinat). Niin ikään epätoivottavia asioita televisiossa olivat Aamulehden ja Turun Sanomien mukaan *väkivalta* ja *pornografia*, Helsingin Sanomien mukaan *televisioväkivalta* ja *televisiopornografia*.

Epäyhdenmukaisuus tiedonhaun näkökulmasta

Kun tiedonhaussa pyritään suureen saantiin eli löytämään mahdollisimman moni relevanteista dokumenteista, tulee hakulausekkeessa käytettävien ilmaisujen vastata mahdollisimman suuressa määrin niitä ilmaisuja, joita haun kannalta relevanteissa dokumenteissa on käytetty. Hakijalla tulisi olla käsitys siitä, millaisia ja millä tavalla erilaisia ilmaisut kyseisestä asiasta kirjoitettaessa voivat eri dokumenteissa olla, jotta hän voisi ottaa erilaiset vaihtoehdot mukaan hakuunsa. Tässä tutkimuksessa tuli esille useitakin epäyhdenmukaisuuden lajeja. Toisiaan vastaavien ilmaisujen väliltä löytyi esimerkiksi erilaisia semanttisia suhteita. Näitä voi olla edustettuna myös haun kohteena

olevissa dokumenteissa, joten pyrittäessä suureen saantiin on hakua yleensä syytä laajentaa lisäämällä hakulausekkeeseen alkuperäiseen hakuavaimeen ekvivalenssi-, hierarkia- ja assosiaatiosuhteessa olevia ilmaisuja. Ilmaisujen ideoinnissa voi käyttää hakutesauruksia, synonyymisanastoja ja muita tiedonhaun apuvälineitä. Joskus myös hakujärjestelmä voi auttaa hakijaa muistuttamalla vaihtoehtoisten ilmaisujen olemassaolosta tai tarjoamalla suoraan ehdotuksia haussa käytettäviksi ilmaisuiksi. (Aitchison & Gilchrist 1987, 119-120; Kristensen 1992, 18; Perez 1982, 189-190; Iivonen 1995, 226-227)

Tutkimuksessa todettiin myös, että eri dokumenttien toisiaan vastaavissa ilmaisuissa saattaa olla yhteisiä osia. Tiedonhakujärjestelmissä on usein ominaisuuksia, jotka helpottavat tällaisten ilmaisujen hakemista. Yksi niistä on hakuavaimen katkaisu, jonka avulla voidaan hakea kätevästi ilmaisuja, joiden yhteinen aines on sanan alussa. Tämä koskee erityisesti johdoksia, joita suomessa yleensä muodostetaan päätteitä sanoihin liittämällä, mutta myös yhdyssanoja, joilla on yhteinen alkuosa. Hakijan ongelmana on vain osata katkaista hakuavain oikeasta kohtaa, jotta kaikki tarvittavat ilmaisut tulevat haun piiriin.

Se, miten haetaan yhdyssanoja, joilla on yhteinen jälkiosa, riippuu siitä, mikä on mahdollista kyseisessä hakujärjestelmässä. Jos yhdyssanat on tallennettu järjestelmään sekä kokonaisina että yhdysosina, voidaan samaloppuisia yhdyssanoja hakea samalla hakuavaimella riippumatta siitä, mikä niiden alkuosa on. Samalla tavalla ovat haettavissa ne käytännössä yleiset tapaukset, joissa sama sana esiintyy dokumenteissa sekä itsenäisenä että yhdyssanan osana. Jos yhdyssanan osilla ei ole mahdollista hakea, jää yleensä ainoaksi keinoksi yrittää keksiä, mitä alkuosia jälkiosiin on yhdistetty ja hakea jokaisella yhdistelmällä erikseen. Tämä voi olla vaikeaa etenkin lehtikielen ollessa kyseessä, sillä siinä käytetään monesti tilapäisiä, vakiintumattomia yhdyssanoja. (Ratkaisuksi hyvin sopiva hakuavaimen katkaisu vasemmalta ei ole käytössä yhtä yleisesti kuin katkaisu oikealta - ks. asiaan liittyvistä ongelmista esim. Alkula & Honkela 1992, 16.)

Tulosten arviointia

Tutkimus vahvisti osaltaan sitä käsitystä, että luonnollisella kielellä kirjoitettujen tekstien sananvalinnassa esiintyy epäyhdenmukaisuutta dokumenttien välillä. Tutkimuksessa tuli esille erilaisia epäyhdenmukaisuuden lajeja. Monesti oli kysymys asioista, jotka muissa yhteyksissä (ks. esim. McKinin et al. 1991, 303) on mainittu ongelmina, jotka liittyvät luonnollisen kielen käyttöön tiedonhaussa. Edellä epäyhdenmukaisuustapauksia tarkasteltiin kahdesta näkökulmasta: millaisia semanttisia suhteita on niiden ilmaisujen välillä, jotka vastaavat sisällöllisesti toisiaan eri lehtien jutuissa ja toisaalta kuinka samankaltaisia tällaiset ilmaisut ovat ulkonaisesti. Näyttää siltä, että epäyhdenmukaisuustapaukset ovat varsin heterogeeninen joukko. Samoin vaihtelevat myös tiedonhaun keinot epäyhdenmukaisuuden haitallisten vaikutusten lieventämiseksi.

Samoista uutisaiheista kirjoitettujen uutisjuttujen sananvalinnasta löytyi epäyhdenmukaisuutta lehtien väliltä. Yhdenmukaisuusarvojen vaihtelu aineiston sisällä oli myös suurta, esimerkiksi sähkeuutiset ja laajat uutiset olivat hyvin erilaisia yhdenmukaisuudeltaan keskiarvoilla mitaten; ryhmien ero oli myös tilastollisesti merkitsevä. Toisaalta samaakin laajuutta ja tyyppiä edustavilla uutisilla saattoi olla hyvin suuria eroja yhdenmukaisuusarvoissa. Aikaisemmissa tutkimuksissa saadut yhdenmukaisuusarvot ovat usein osoittautuneet varsin alhaisiksi, ja ne ovat myös vaihdelleet paljon tutkimuksesta toiseen. Arvojen suora vertailu on kuitenkin kyseenalaista, koska mm. yhdenmukaisuuden laskutavat ovat saattaneet vaihdella tutkimuksesta toiseen. (Iivonen 1993, 64-65) Silloin kun käsitteiden yhdenmukaisuutta on tutkittu, on se poikkeuksetta osoittautunut suuremmaksi kuin ilmaisujen yhdenmukaisuus (emt.). Näin oli myös tässä tutkimuksessa: uutisjutut olivat huomattavasti yhdenmukaisempia käsitteiltään kuin ilmaisuiltaan.

Mitä tutkimuksessa saadut yhdenmukaisuusarvot sitten kertovat ? Ne kertovat siitä, missä määrin vertailluissa uutisjutuissa oli samoja (keskeisiä) käsitteitä ja (keskeisiä käsitteitä vastaavia) ilmaisuja. Niistä ei voi kuitenkaan suoraan päätellä, mikä vaikutus tilanteella olisi tiedonhakuun. Yleisellä tasolla voidaan todeta, että epäyhdenmukaisuudella on negatiivinen vaikutus tiedonhakuun: hakijan tulee ottaa vaihtoehtoisia ilmaisuja hakuunsa, ja osa relevanteista dokumenteista voi jäädä löytymättä, jos vaihtoehtojen ideoinnissa ei onnistuta. Kuitenkaan kaikki epäyhdenmukaisuus ei ole yhtä haitallista tiedonhaun kannalta. Esimerkiksi kun yhdessä lehdessä käytettiin alppihiihtäjä Janne Leskisestä nimen ja sanan *alppihiihtäjä* lisäksi ilmaisuja *hurjapää, ässä* ja *taistelija*, ei liene kovin huolestuttavaa tiedonhaun näkökulmasta, että kolme viimeksi mainittua ilmaisua puuttuivat kahdesta muusta lehdestä.

Tässä tutkimuksessa kaikkia ilmaisuja ja käsitteitä pidettiin samanarvoisina ottamatta mitenkään

huomioon niiden todennäköistä käyttöarvoa tiedonhaussa. Näin ollen tämä tutkimus saattaakin antaa hieman liian synkän kuvan tilanteesta, varsinkin kun tiedonhaun kannalta marginaalinen aines helposti on keskimääräistä epäyhdenmukaisempaa. Jotta tilanteesta saataisiin todellisuutta paremmin vastaava kuva, tulisi tutkimus rajata ensisijaisesti niihin ilmaisuihin, joilla voitaisiin ajatella olevan käyttöä hakuavaimina tiedonhaussa. Tätä varten tulisi kehittää kriteerejä, joilla voitaisiin arvioida ilmaisujen tiedonhakuarvoa. Kun tämä oli tutkimus dokumenttien ehdoilla, olisi tulevassa tutkimuksessa mielestäni syytä ottaa paremmin huomioon tiedonhaku ja sen realiteetit.

Hyväksytty julkaistavaksi 15.3.2000.

Lähteet

Aitchison, Jean & Gilchrist, Alan (1987). Thesaurus construction: a practical manual. London: Aslib.

Alkula, Riitta & Honkela, Timo (1992). Tekstin tallennus- ja hakumenetelmien kehittäminen suomen kielen tulkintaohjelmien avulla. Espoo: VTT.

Gale directory of databases (1997). Vol. 1: Online databases. Detroit (Mich.): Gale Research.

Häkkinen, Kaisa (1990). Mistä sanat tulevat: suomalaista etymologiaa. Helsinki: Suomalaisen kirjallisuuden seura.

Iivonen, Mirja (1993). Johdonmukaisuuden laskeminen tiedon tallennuksen ja haun tutkimuksessa. - Kirjastotiede ja informatiikka 12(2): 63-76.

Iivonen, Mirja (1995). Hakulausekkeiden muotoilun yhdenmukaisuus onlineviitehaussa. Tampere: Tampereen yliopisto.

ISO (1986). ISO International Standard 2788. International Organization for Standardization.

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

Karlsson, Fred (1994). Yleinen kielitiede. Helsinki: Yliopistopaino.

Kristensen, Jaana (1992). Vapaatekstihakujen laajentaminen hakutesauruksen avulla haettaessa indeksoimattomasta tekstitietokannasta. Tampereen yliopisto: Kirjastotieteen ja informatiikan lisen-siaattitutkielma.

Lehtokangas, Raija (1999). Sananvalinnan yhdenmukaisuus sanomalehtiuutisissa. Tampereen yliopisto: Informaatiotutkimuksen pro gradu -tutkielma.

Lepäsmaa, Anna-Liisa & Lieko, Anneli & Silfverberg, Leena (1996). Miten sanoja johdetaan: suomen kielen johto-oppia. Helsinki: Finn Lectura.

McKinin, Emma Jean & Sievert, MaryEllen & Johnson, E. Diane & Mitchell, Joyce A. (1991). The Medline / Full-text research project. - Journal of the American Society for Information Science 42(4): 297-307.

Miettinen, Jorma (1984). Toimitustyö: journalistiksi suunnistautuvan oppikirja. Helsinki: Gaudeamus.

Niemi, Minna (1997). Sanomalehtiartikkelien sisällön kategoriat mahdollisina tiedonhaun apuvälineinä. Tampereen yliopisto: Informaatiotutkimuksen sivulaudaturtutkielma.

Okkonen, Antero (1986). Toimittajan työ 1. Hämeenlinna: Karisto.

Olkinuora, Katri (1992). Onko aalto suurempi kuin laine ? Synonymian ongelmia. - Variaatioita: Tampereen yliopiston suomen kielen ja yleisen kielitieteen laitoksen juhlakirja, s. 113-124. Tampere: Tampereen yliopisto.

Perez, Ernest (1982). Text enhancement: controlled vocabulary vs. free text. - Special libraries 73(3): 183-192.

Sivula, Jaakko (1989). Synonyymit. - Nykysuomen sanavarat (toim. Jouko Vesikansa), s. 183-199. Helsinki: WSOY.

Vesikansa, Jouko (1978). Johdokset. Helsinki: WSOY.