Erkka Leppänen

Homografiongelma tekstihaussa ja homografien disambiguoinnin vaikutukset

Leppänen, Erkka, Homografiongelma tekstihaussa ja homografien disambiguoinnin vaikutukset [The homonymy problem in free-text searching and the results of the homonymy disambiguation]. Informaatiotutkimus 15 (4): 133–144, 1996.

Homonymy is known to often cause false drops in free-text searching in a full-text database. The problem is quite common and difficult to avoid in Finnish but nobody has examined it before. This article is based on a study that examined the frequency of and solutions to the homonymy problem. Searches were made in a Finnish full-text database containing about 55 000 newspaper articles. The results indicate that homonomy is not a very serious problem in free-text searching. Only about one search result set out of four contained false drops caused by homonymy. Several other reasons of irrelevance were much more common. However, in some result sets there was a considerable number of homonymy errors, so the problem seems to be very random. It was also studied whether homonymes can be disambiguated by a syntactic analysis. The result is that 75.2 % of homonymes could be disambiguated by this method. Verb homonymes were considerably easier to disambiguate than substantive homonymes. Although homonymy is not a very big problem it perhaps could be easily eliminated if there was a syntactic analyzer component in the IR system.

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

Johdanto

Hakuvirheet ovat tiedonhaun arkipäivää. Sataprosenttisesti onnistuneisiin hakuihin päästään tuskin koskaan, ja etenkin laajoista tekstitietokannoista on vaikea tehdä tarkkoja hakuja. Vaikeudet voivat olla käytetystä kielestä riippumattomia, mutta usein ne ovat kielisidonnaisia. Haettaessa suomenkielisestä tekstikannasta kohdataan hankaluuksia, joita ei tarvitse ottaa huomioon englanninkielisessä tietokannassa, ja sama pätee myös toisin päin.

Yksi suomen kielen ongelmista on monitulkintaisten sanamuotojen – homonyymien – suuri määrä. Noin 15 % suomen kielen sanoista on homonyymisia (Karlsson 1994, 80). Homonyymeja on muissakin kielissä, mutta suomessa niitä on erityisen runsasmääräisesti sanojen taipumisen vuoksi. Ne heikentävät hakujen tarkkuutta, ja lisäksi ne aiheuttavat pulmia suomen kieltä käsitteleville tietokoneohjelmille.

Homonymialla tarkoitetaan kahden tai useamman eri sanan rakenteellista monitulkintaisuutta (Karlsson 1994, 197). Homonyymit voidaan jaotella sen mukaan, onko

homonymia täydellinen vai osittainen. Homonymia on täydellinen, kun sanan kaikki taivutusmuodot ovat samanasuisia (esim. *halli*). Lähellä täydellisiä homonyymeja ovat sanat, joiden perusmuodot ovat samanasuisia mutta joiden taivutusmuodoista osa on keskenään eriasuisia (esim. *kuusi*). Osahomonyymit eli sanamuotohomonyymit ovat samanmuotoisia yleensä vain yhdessä taivutusmuodossaan (esim. *hukkaan, alusta*). (Laalo 1990, 27–28.)

Jos halutaan korostaa, että homonyymi esiintyy samanmuotoisena nimenomaan kirjoitetussa kielessä, voidaan käyttää käsitettä ***homografi***. Vastaavasti puhutussa kielessä esiintyvää homonyymia voidaan kutsua ***homofoniksi***. Suurin osa suomen homonyymeista on kuitenkin sekä homografeja että homofoneja. (Laalo 1990, 34–35.) Kirjoitetussa tekstissä esiintyviä sanamuotohomonyymeja kutsutaan tässä tutkimuksessa ***osahomografeiksi***.

Useimmiten tottunut suomen kielen käyttäjä ei edes huomaa homografisia sanamuotoja, sillä tekstiyhteys ohjaa automaattisesti oikeaan tulkintaan. Joskus oikea tulkinta vaatii kuitenkin jonkin verran pohtimista, sillä homografit voivat esiintyä myös siten, ettei niiden todellista merkitystä pysty selvittämään, ellei tunne laajempaa kontekstia. (Laalo 1990, 12–13.) Monet sanaleikit perustuvat juuri ällaiseen monitulkintaisuuteen:

> Kun ostaa maastoauton, ei tarvitse *katua*.
> Tekisi mieli *kiljua*.
> Paljonko maitoa saadaan kymmenestä lehmästä päivässä, kun *kustakin* saadaan kahdeksan litraa?

Eräät homografimuodot ovat niin harvinaisia, että niitä käytetään tuskin koskaan normaalissa kielessä. Esimerkiksi sanamuoto *kuin* on paitsi yleinen konjunktio myös kuusubstantiivin monikon instruktiivi, mutta on vaikea keksiä sille käyttöä *kuu*-merkityksessä. Runoilija voisi tosin ehkä sepitellä säkeen *"yötaivas hohtaa yllämme tähdin ja kuin"* edellyttäen, että runoilijan kotiplaneetalla on enemmän kuin yksi kuu.

Myös erisnimet voivat olla homografisia. Tavallisesti pronominina esiintyvä sana *se* voi tarkoittaa myös samannimistä rockyhtyettä tai lehteä. Vastaavasti sanat *valio* ja *neste* voivat tarkoittaa yrityksiä. *Kotka* voi tarkoittaa lintulajia tai kaupunkia. *Ankara* voi tarkoittaa luonteenpiirrettä kuvaavaa adjektiivia, Turkin pääkaupunkia tai Tuomari Nurmion samannimistä iskelmää. Homonyymien todellisen merkityksen ratkomista kutsutaan ***disambiguoinniksi***.

Tekstihaussa homografit aiheuttavat ongelmia vain siinä tapauksessa, että tiedonhakujärjestelmälle syötettävien hakusanojen joukossa on homografeja. Kun tiedonhakujärjestelmä suorittaa hakua, se ei kykene erottamaan, missä merkityksessä homografisana esiintyy dokumentissa. Niinpä tulosjoukkoon voi päästä mukaan sellaisiakin dokumentteja, joissa kyseinen sana esiintyy ainoastaan "väärässä" merkityksessä. (Sormunen & Alkula 1990, 32.)

Suomen tuhansista homografisista sanamuodoista läheskään kaikki eivät aiheuta ongelmia tiedonhaussa. Monet homografisanat ovat niin yleismerkityksellisiä, että on vaikea kuvitella hakukysymystä, jossa niitä tarvittaisiin hakusanoina (esim. *eli, kumman, vähän*). Joissain tapauksissa homografiasta saattaa olla jopa hyötyä. Näin voi käydä silloin, kun homografin kaksi eri merkitystä ovat semanttisesti lähellä toisiaan. Esimerkiksi sanamuoto *lainaa* voi tarkoittaa sekä substantiivia *laina* että verbiä *lainata*. Jos tiedonhakija hakee tietoa lainaamiseen liittyvästä aiheesta ja sattuu unohtamaan joko *laina-* tai *lainata*-hakusanan, homografinen hakusana voi tuoda dokumentteja, joissa sanamuoto lainaa esiintyy eri sanan taivutusmuotona kuin millä alunperin haettiin mutta jotka silti ovat relevantteja. Ongelmia sen sijaan saattaa aiheuttaa se, että *Laina* on myös naisennimi.

Useissa aikaisemmissa suomenkielisten tekstikantojen tutkimuksissa on jo havaittu homografiongelman olemassaolo. Kristensenin tutkimuksessa 1989 homografit luokiteltiin yhdeksi kuudesta virhetyypistä, jotka heikensivät hakujen tarkkuutta. Kuitenkin vain 3,8 % epärelevanteista artikkeleista oli epärelevantteja homografian takia. Useimmat muut virhetyypit olivat huomattavasti yleisempiä. Homografiongelman todettiin kuitenkin olevan vaikeasti vältettävissä. Eniten ongelmia aiheuttivat sanamuodot *lailla* ja *lainkaan*. (Kristensen & Järvelin 1990, 81.)

Myös Riitta Alkulan ja Timo Honkelan FULLTEXT-projektissa törmättiin homografiongelmaan. Tarkkuusvirheitä aiheuttivat ainakin sanamuodot *Halvan*, (Leif) *Salmén* ja (Inga) *Sulin* (Alkula & Honkela 1992, 84–85, 88–89). Tosin näitä hakusanoja käytettiin hauissa nimenomaan siksi, jotta saataisiin selville ongelmallisten sanojen käyttäytyminen erilaisissa tiedonhakujärjestelmissä, eivätkä ne siksi kerro mitään homografiongelman yleisyydestä.

Eero Sormusen lisensiaattitutkimuksessa (1994) kohdattiin niinikään homografivirheitä, mutta niistä ei raportoitu erikseen. Kyseisen tutkimuksen todellinen homografivirheiden määrä selviää tässä tutkimuksessa, sillä tämä perustuu suurelta osin samaan aineistoon.

Tämän tutkimuksen tavoitteena oli selvittää, paranisiko hakujen tarkkuus oleellisesti ja katoaisivatko homografiasta johtuvat hakuvirheet, jos tiedonhakujärjestelmä pystyisi disambiguoimaan homografit. Samalla tutkittiin, minkä tyyppisiä homografeja on olemassa ja ovatko toiset homografityypit vaikeammin disambiguoitavissa kuin toiset. Huomiota kiinnitettiin erityisesti homografien sanaluokkaan sekä lauseenjäsenyyteen. Lopuksi esitettiin tutkimuksessa ilmitulleisiin seikkoihin pohjautuva malli tietokoneohjelmalle, joka suorittaisi disambiguoinnin. (Leppänen 1995.)

Tutkimuksen lähtökohdat

Aikaisemmat tekstihaun tutkimukset sekä käytännön työ tekstikantojen parissa ovat osoittaneet, että homografit aiheuttavat ongelmia ja heikentävät hakutuloksia. Ongelman olemassaolo oli siis todettu jo ennen tätä tutkimusta, mutta sen laajuutta ja ratkaisumahdollisuuksia ei ollut tutkittu sen tarkemmin.

Tämän tutkimuksen ensisijainen tavoite oli selvittää, kuinka paljon tarkkuusvirheitä homografit aiheuttavat tekstihaussa. Tutkimusaineistona käytettiin samoja hakuja ja tulosjoukkoja kuin Eero Sormusen lisensiaattitutkimuksessa (1994). Tosin lopulliseen tutkimukseen pääsivät vain ne haut, joiden tulosjoukoissa todella oli homografivirheitä. Kun homografivirheet oli tunnistettu, pystyttiin kertomaan, miten paljon hakutulokset olisivat parantuneet, jos homografivirheitä ei olisi ollut.

Toinen tärkeä tavoite oli pohtia, miten homografien disambiguoinnin voisi parhaiten toteuttaa. Tätä tutkimusvaihetta varten tehtiin aivan uusia hakuja käyttäen apuna Suomen kielen homonyymiluetteloa (Saukkonen ym. 1982). Hakutulosten perusteella pystyttiin päättelemään, minkä tyyppiset homografit ovat hankalimpia ja mitä ongelmia niiden disambiguoinnissa olisi. Homografien disambiguoinnin ajateltiin perustuvan lähinnä juoksevan tekstin syntaktiseen analyysiin (= lauseenjäsennykseen), mutta muutkin mahdollisuudet pidettiin mielessä.

Haut tehtiin tutkimustietokannasta, joka sisältää noin 55 000 Aamulehdessä, Keskisuomalaisessa ja Kauppalehdessä vv. 1990–92 ilmestynyttä sanomalehtiartikkelia. Tietokantaa kontrolloi *Topic*-tiedonhakuohjelma, jonka hakuominaisuudet perustuvat Boolen operaattoreiden käytölle. Topic ei pysty disambiguoimaan homografeja, joten tutkimus oli tehtävä mallintaen eli päätellen tuloksista, millainen vaikutus disambiguoinnilla olisi ollut.

Homografiongelman kartoittaminen

Tulosten arviointiperusteet

Ensimmäisen vaiheen tutkimusaineisto koostui valmiista hakukysymyskokoelmasta ja tulosjoukoista, jotka olivat peräisin Eero Sormusen lisensiaattityöstä (1994). Tämän tutkimusvaiheen päätavoitteena oli kartoittaa homografivirheiden yleisyyttä. Aineisto oli koottu ja analysoitu etukäteen, mutta homografivirheiden määrää siinä ei ollut aiemmin selvitetty.

Tiedonhaun tuloksien arvioimiseen käytetään yleensä käsitteitä ***saanti*** ja ***tarkkuus***. Tarkkuus ilmaisee, montako prosenttia tulosjoukon dokumenteista on relevantteja. Saanti taas ilmaisee, montako prosenttia kaikista tietokannan sisältämistä relevanteista dokumenteista on tulosjoukossa. Lähes poikkeuk-

setta pyrkimys hyvään saantiin laskee tarkkuutta ja päinvastoin. Tämän tutkimuksen kannalta tarkkuus on saantia tärkeämpi käsite, koska homografien disambiguoinnin pitäisi vaikuttaa nimenomaan tarkkuuteen. Saantia se ei todennäköisesti muuttaisi lainkaan, mutta se voisi parantaa tarkkuutta karsimalla tulosjoukosta epärelevantteja dokumentteja.

Relevanssiasteikko oli kaksiportainen: relevantti – epärelevantti. Aiheeltaan marginaaliset artikkelit luokiteltiin epärelevanteiksi. Useimmista tiedonhakututkimuksista poiketen tässä tutkimuksessa kiinnitettiin päähuomio epärelevantteihin artikkeleihin eikä relevantteihin. Epärelevanteista artikkeleista pyrittiin löytämään syy, miksi se oli tullut tulosjoukkoon. Erilaisia virhetyyppejä tunnistettiin kahdeksan:

1) Hakusanojen välillä ei ole mitään keskinäistä suhdetta tekstissä.
2) Hakusanojen suhde tekstissä on vääräntyyppinen tai virheellinen hakukysymyksen kannalta.
Esimerkiksi hakukysymyksessä haetaan tietoja pääministeri Margaret Thatcherin erottamisesta ja käytetään hakusanoja *Thatcher* ja *erottaa*. Haussa löydetäänkin artikkeli, jossa kerrotaan Thatcherin erottaneen jonkun ministerinsä.
3) Homografian aiheuttamat virheet.
4) Hakuaihetta käsitellään tekstissä niin marginaalisesti, ettei artikkelia voida pitää relevanttina.
5) Hakusanojen katkaisusta johtuvat virheet.
6) Polysemian (= hakusanan semanttisen monimerkityksellisyyden) aiheuttamat virheet.
7) Tiedonhakujärjestelmän ominaisuuksien aiheuttamat virheet.
8) Painovirheistä johtuvat virheet.

Tämän tutkimuksen puitteissa ei ollut resursseja tehdä tarkkaa hakuvirheanalyysia, vaan päähuomio kiinnitettiin virhetyyppiin 3). Yleisimmät virhetyypit vaikuttivat kuitenkin olevan tyypit 1), 2), 4) ja 5). Tosin samankin virhetyypin frekvenssit saattoivat olla hyvin erilaisia eri tulosjoukoissa.

Kaikkien epärelevanttien artikkeleiden kohdalla virhetyyppiä ei pystynyt yksiselitteisesti nimeämään, sillä eräät niistä olisi voinut perustellusti luokitella vähintään kahteen virhetyyppiin. Vaikka yhden virheen olisi eliminoinut, olisi toisentyyppinen virhe pitänyt artikkelin edelleen tulosjoukossa. Homonyymivirheartikkeleiksi määriteltiin loppujen lopuksi vain sellaiset artikkelit, jotka olisivat karsiutuneet tulosjoukosta disambiguoimalla homografit.

Tulosten perusteella pyrittiin myös arvioimaan, onko hakukysymyksen ominaisuuksilla vaikutusta homografivirheiden määrään. Tutkitut ominaisuudet olivat hakukysymyksen *käsitetyyppien luonne* (yksilökäsitteet ja yleiskäsitteet), *kompleksisuus* (rajaavien käsitteiden määrä) sekä *laaja-alaisuus* (hakusanojen määrä rajaavaa käsitettä kohti).

Taulukko 1. Hakukysymykset, joiden tulosjoukoista löytyi homografivirheitä. Hakujen numerointi perustuu Eero Sormusen lisensiaattityössä käytettyyn numerointiin.

Haun numero	Hakukysymys
3	Suomen metsäteollisuuden polkumyyntisyytökset USA:ssa.
4	Jyväskylän kaupungin ja maalaiskunnan kuntaliitoshanke.
13	Carl Bildtin lausunnot Suomen ja Ruotsin yhteistyöstä.
18	Mitä tahansa taustatietoja Valion toiminnasta.
21	Keran ja KTM:n investoinnit matkailuyrityksiln.
22	Neste Oy:n maakaasutoiminta.
25	Elintarvikkeiden tuontirajoitusten poisto Suomessa.
35	Vihreiden kansanedustajien Suomen eduskunnassa tekemät aloitteet.

Taulukko 2. Sanamuodot, jotka aiheuttivat homografivirheitä eri hauissa.

Haun numero	Homografivirheitä aiheuttaneet sanamuodot
3	*kymmeneen, kymmenen, kymmenessä, kymmenestä*
4	*liitosta*
13	*suomi*
18	*valio, valiona, voi, voimme, voin, voisi*
21	*kera*
22	*neste, nesteensä*
25	*tuo, tuon*
35	*vihreillä, vihreitä, vihreä, vihreällä, vihreän, vihreässä, vihreätä, vihreää, vihreään*

Tutkimustulokset

Aineistossa oli kaikkiaan 35 hakukysymystä ja tulosjoukkoa. Mahdollisuus homografivirheisiin havaittiin kuitenkin vain 15 hakukysymyksessä eli 42,9 prosentissa, sillä vain näiden hakujen hakulausekkeissa oli homografisia hakusanoja. Kun tulosjoukot käytiin tarkemmin läpi, osoittautui, että homografivirheitä oli vain kahdeksassa tulosjoukossa näistä viidestätoista. Kaikkiaan homografivirheitä oli siis 22,9 prosentissa kaikista tulosjoukoista eli lähes joka neljännessä. Kahdeksan homografivirheitä aiheuttanutta hakukysymystä on esitelty taulukossa 1. Taulukossa 2 on lueteltu sanamuodot, jotka aiheuttivat homografivirheitä.

Useimmista kahdeksasta hakukysymyksestä tehtiin kompleksisempia alihakuja, ja hakujen yhteismääräksi tuli lopulta 19. Näin tulosjoukkojen yhteismäärä oli myös 19, mutta 11 niistä oli vain suurempien tulosjoukkojen osajoukkoja.

Näissä 19 tulosjoukossa homografivirheiden osuus kaikista hakuvirheistä vaihteli suuresti. Pienimmillään se oli vain 0,7 prosenttia, suurimmillaan 45,7 prosenttia. Keskimäärin se oli 16,0 prosenttia. Enimmillään homografien disambiguointi olisi nostanut

Taulukko 3. Hakukysymyksen ominaisuudet eri hauissa sekä hakujen tarkkuus.

Haun numero	Hakutyyppi	Kompleksisuus	Laaja-alaisuus	Tarkkuus	Tarkkuus ilman homografivirheitä + *(suht. parannus)*	
3	Yleis	2	15,5	27,1 %	40,6 %	*(49,8 %)*
4	Yleis	4	7,5	5,2 %	5,2 %	*(0,0 %)*
13	Yksilö	3	2,0	38,9 %	43,8 %	*(12,6 %)*
18	Yksilö	2	20,0	15,0 %	16,9 %	*(12,7 %)*
21	Yksilö	2	21,0	20,6 %	23,6 %	*(14,6 %)*
22	Yksilö	2	5,5	26,3 %	26,8 %	*(1,9 %)*
25	Yleis	3	24,0	12,0 %	12,1 %	*(0,8 %)*
35	Yksilö	2	2,0	12,0 %	14,3 %	*(19,2 %)*
Keskimäärin		2,5	12,2	19,6 %	22,9 %	*(14,0 %)*

haun tarkkuutta 27,1:stä 40,6 prosenttiin eli 49,8 prosenttia. Sen sijaan eräissä hauissa tarkkuus ei olisi parantunut edes promillea, vaikka homografit olisi disambiguoitu.

Kahdeksan päähaun keskimääräinen tarkkuus oli 19,6 prosenttia. Jos kaikki homografit olisi pystynyt disambiguoimaan, olisi hakujen keskimääräinen tarkkuus ollut 22,9 prosenttia ja parannusta olisi tullut keskimäärin 14,0 prosenttia. Täytyy kuitenkin muistaa, että nämä luvut koskevat vain niitä hakuja, joissa homografivirheitä ylipäänsä oli. Jokaista homografivirheitä sisältävää tulosjoukkoa kohden oli yli kolme tulosjoukkoa, joissa homografivirheitä ei ollut lainkaan.

Hakukysymysten sekä tulosjoukkojen ominaisuudet on esitelty taulukoissa 3 ja 4. Yleis- ja yksilökäsitehakujen välillä ei päähauissa havaittu kovin selkeää eroa homografivirheiden suhteen. Yleiskäsitehauissa homografivirheiden osuus kaikista hakuvirheistä oli keskimäärin 15,9 %. Yksilökäsitehauissa vastaava luku oli 13,5 %. Yleiskäsitehauissa homografien disambiguointi olisi parantanut keskimääräistä tarkkuutta 16,9 %, yksilökäsitehauissa hieman vähemmän eli 12,2 %. Hyvin usein homografivirheitä aiheutti juuri yksilökäsitettä kuvaava hakusana (esim. *Neste, Kera*). Kaikki haut, joissa on yksilökäsitteitä hakusanoina, eivät kuitenkaan välttämättä ole yksilökäsitehakuja.

Sen sijaan haun laaja-alaisuudella tuntui olevan jonkin verran vaikutusta homografivirheiden määrään. Niissä päähauissa, joissa haun laaja-alaisuus oli yli keskiarvon (12,2) homografivirheiden osuus kaikista hakuvirheistä oli 19,1 prosenttia. Kapea-alaisemmissa hauissa homografivirheiden osuus oli vain 9,7 prosenttia. Vaikuttaisi siis siltä, että mitä enemmän hakulausekkeessa on hakusanoja, sitä enemmän tulosjoukossa on homografivirheitä, mikä on aivan loogista. Ero on kuitenkin niin pieni, että kyse saattaa olla pelkästä sattumastakin.

Hakukysymyksen kompleksisuuden ja homografivirheiden välinen suhde ei sekään ole täysin selvä. Selvää on vain, että homografivirheiden lukumäärä ei voi lisääntyä kompleksisuuden kasvaessa, vaan se voi ai-

Taulukko 4. Tulosjoukkojen koot sekä homografivirheiden määrä ja osuus eri hauissa.

Haun numero	Tulosjoukon koko	Homografi-virheitä	Homografivirheiden osuus hakuvirheistä
3	48	16	45,7 %
4	154	1	0,7 %
13	18	2	18,2 %
18	280	32	13,4 %
21	63	8	16,0 %
22	129	2	2,1 %
25	184	2	1,2 %
35	108	17	17,9 %
Keski-määrin	123,0	10,0	14,4 %

noastaan vähentyä. Sen sijaan niiden suhteellinen osuus saattaa jopa lisääntyä. Rajaavien käsitteiden lisääminen ei ole niin hyvä keino homografivirheiden torjumiseksi, kuin ehkä voisi luulla.

Toisaalta, kun muistetaan, että peräti 15 alkuperäisistä tulosjoukoista saattoi hakulausekkeidensa perusteella sisältää homografivirheitä mutta vain kahdeksan todella sisälsi niitä, huomataan, että suuri osa homografeista karsiutuu jo siinä vaiheessa kun ensimmäinen rajaava käsite lisätään hakukysymykseen. Tämän perusteella voidaan sittenkin uskoa, että suurin osa homografivirheistä katoaa tulosjoukosta hakukysymyksen kompleksisuuden kasvaessa. Jäljellejäävät voivat kuitenkin olla sitäkin vaikeammin vältettävissä.

Pelkän hakukysymyksen perusteella ei siis ole helppo ennustaa, tuleeko tulosjoukkoon homografivirheitä. Kyse tuntuu olevan hyvin sattumanvaraisesta ongelmasta. Yleensä kuitenkin ongelmia voidaan odottaa, jos jokin haun keskeisimmistä käsitteistä on homografinen jonkun yleisen sanan kanssa.

Loppujen lopuksi homografit aiheuttivat huomattavasti hakuvirheitä ainoastaan yhdessä päähaussa 35:stä eli haussa 3. Useimmissa muissa hauissa homografiongelma oli marginaalinen tai jopa olematon verrattuna muihin hakuvirheitä aiheuttaviin tekijöihin. Johtopäätöksenä voidaan siis sanoa, että homografit tuskin ovat se ongelma, johon tiedonhaun kehittäjien kannattaisi ensisijaisesti kiinnittää huomiota.

Homografivirheiden kartoittamisen lisäksi tässä tutkimusvaiheessa pohdittiin keinoja homografiongelman ratkaisemiseksi. Eräs merkittävä havainto oli, että suuri osa homografivirheistä poistuisi, jos erisnimet voitaisiin tunnistaa erisnimiksi esimerkiksi ison alkukirjaimen perusteella. Tällöin esimerkiksi sellaiset hakusanat kuin *Kymmene* ja *Valio* eivät toisi tulosjoukkoon homografivirheitä.

Toinen havainto koski sanomalehtitekstin otsikoita. Otsikot eivät yleensä ole kieliopin mukaisia lauseita, ja näin ollen niihin ei tehoaisi samanlainen syntaktinen analyysi kuin muuhun tekstiin. Esimerkiksi otsikossa *"Armenian herkkuja kansanmusiikin **kera**"* ei ole lainkaan predikaattia. Vastaavanlaisia epätäydellisiä lauseita esiintyy varmasti myös muuntyyppisessä tekstissä.

Homografien disambiguoituvuus

Tulosten arviointiperusteet

Tämän vaiheen tutkimusaineisto koostui homografisista hakusanoista sekä niillä saaduista tulosjoukoista. Tavoitteena oli tutkia erityyppisten homografien disambiguoinnin mahdollisuuksia. Aineisto koottiin samasta tutkimustietokannasta, jota käytettiin edellisessäkin tutkimusvaiheessa.

Homografien tyypittely perustui Suomen kielen homonyymeja -luettelossa (Saukkonen ym. 1982) käytettyyn jaotteluun, jossa homografityypit luokitellaan sen mukaan, mihin sanaluokkiin kunkin homografin eri perusmuodot kuuluvat. Esimerkiksi homografin *niitä* perusmuoto on joko verbi *niittää* tai pronomini *ne*, joten se luokitellaan verbien ja pronominien kombinaatioon. Hakusanoina käytettiin erilaisia homografeja. Päähuomio kiinnitettiin substantiivien ja verbien homografiaan, koska näitä sanaluokkia käytetään tiedonhaussa ylivoimaisesti eniten.

Tulosjoukkoja tutkittiin siten, että tutkija yritti disambiguoida kutakin homografia mallintamalla syntaktisen analyysin toimintaa. Erityisesti kiinnitettiin huomiota sellaisiin homografeihin, jotka eivät olisi yhtä yksiselitteisiä tietokoneelle kuin ihmiselle. Tämä kuvitteellinen syntaktinen analyysi, jota tässä tutkimuksessa käytettiin mallintamaan tietokoneen suorittamaa lauseenjäsennystä, perustuu mallille, jossa tietokone joutuu ensin jäsentämään lauseen ennen kuin se pystyy löytämään homografisille sanoille oikean sanaluokan ja merkityksen.

Tutkimustulokset

Tutkittavia homonyymityyppejä oli kaikkiaan kolmekymmentä. Näistä pyrittiin valitsemaan hakusanoiksi sellaisia sanoja, joita voisi ajatella käytettävän tiedonhaussa. Kustakin homonyymityypistä valittiin korkein-

taan kymmenen homografin otos. Hakusanoja keksittiin myös luettelon ulkopuolelta. Kutakin homonyymityyppiä kohden hakusanoja oli keskimäärin 5,3. Hakusanojen ja siten myös tehtyjen hakujen yhteismäärä oli 159. Eri tulosjoukkoja oli siis myös 159.

Jokaisesta tulosjoukosta valittiin satunnaisesti kymmenen tutkittavaa artikkelia. Jos artikkeleita oli vähemmän kuin kymmenen, tulosjoukko otettiin tutkimukseen mukaan kokonaisuudessaan. Otoksissa oli yhteensä 1473 artikkelia.

Kun disambiguointi mallinnettiin syntaktista analyysia käyttäen, tutkituissa 1473 artikkelissa 75,2 prosenttia homografeista olisi disambiguoitunut tällä keinolla.

Substantiivit vaikuttivat olevan ongelmallisempia disambiguoitavia kuin verbit. Niiden disambiguoituvuus olisi ollut n. 68,8 %. Substantiivit esiintyvät tavallisesti useampana lauseenjäsenenä kuin verbit, ja näin ollen ne voivat tulla sekoitetuksi muiden sanaluokkien sanojen kanssa. Ehkä kaikkein ongelmallisin kombinaatio oli substantiivien ja adverbien homografiat. Esimerkiksi lauseessa

"Vaurioita ei korjattu *ajoissa*."

homografi on lauseenjäsenyydeltään aina adverbiaali, olipa sen perusmuoto substantiivi *ajo* tai adverbi *ajoissa*. Pelkkä syntaktinen analyysi ei riittäisi disambiguointiin.

Ongelmaa mahdollisesti helpottaisi, jos disambiguointiohjelma kykenisi tunnistamaan yleisimmät fraasit, joissa tietyt adverbit esiintyvät. Jos homografi esiintyy tietyn sanan – yleensä verbin – kanssa, se melkein varmasti on adverbi. Tällaisia fraaseja ovat esim. *käydä toimeen, panna liikkeelle, panna vireille, katsoa perään, ottaa todesta, varteen otettava* ja *jäädä jälkeen*.

Verbit olisi siis helpompi disambiguoida, sillä tämän tutkimuksen perusteella 85,9 % niistä olisi voinut disambiguoida syntaktisen analyysin avulla. Verbit esiintyvät lauseissa useimmiten predikaatteina, ja niitä ei ole helppo sekoittaa muihin lauseenjäseniin. Ellei lauseenjäsennin pysty tunnistamaan predikaattia, ei automaattisessa lauseenjäsennyksessä päästä alkuunkaan. Toisinaan homografimuotoiset verbit kuitenkin esiintyvät muinakin lauseenjäseninä, ja silloin niiden disambiguointi voi olla vaikeampaa, esim:

"Olen väsynyt tähän *toistamiseen*."

Tuntematta laajempaa kontekstia ei voi olla varma, esiintyykö homografi lauseessa *toistaa*-verbinä vai *toistamiseen*-adverbina.

Täydellisiin homografeihin syntaktinen analyysi ei tehoaisi. On myös sellaisia osahomografeja, joita ei pystyisi disambiguoimaan syntaktisin keinoin. Tällaisia ovat ne nominit, jotka ovat homografisia esiintyessään samassa sijamuodossa (*patoihin, hauissa*), sekä ne verbit, jotka ovat homografisia samassa persoona- ja aikamuodossa (*tavata, ammu*).

Verbihomografeihin ehkä tehoaisi kuitenkin toinen keino: verbien paikkaisuuden määrittely. Verbit voivat olla joko nolla-, yksi-, kaksi- tai kolmipaikkaisia sen mukaan, montako pakollista lauseenjäsentä niiden ympärillä on. Esimerkiksi verbi *kuolla* on yksipaikkainen, sillä se vaatii aina subjektin (esim. *Kalle kuoli.*) Kaksipaikkainen verbi vaatii sekä subjektin että objektin tai adverbiaalin. Kolmipaikkainen verbi vaatii kolme lauseenjäsentä (esim. *antaa – Eeva antoi omenan Aatamille.*) Nollapaikkaisia verbejä ovat mm. *sataa* ja *tuulla*. (Karlsson 1994, 145.)

Syntaktinen analyysi ei disambiguoisi esimerkiksi seuraavan lauseen homografia, mutta verbin paikkaisuuden määrittely tehoaisi:

"Puutarhaan *kylvettiin* penkillinen unikoita."

Verbit *kylpeä* ja *kylvää* ovat homografisia eräissä passiivimuodoissaan. Näistä verbeistä *kylpeä* on normaalisti yksipaikkainen ja *kylvää* kaksipaikkainen, mutta passiivimuodossa molempien paikkaluku putoaa yhdellä. Lauseesta huomataan, että predikaatti *kylvettiin* saa objektin *unikoita*. Niinpä kyse ei voi olla *kylpeä*-verbistä, ja ainoaksi vaihtoehdoksi jää *kylvää*.

Verbien paikkaisuuden käyttö disambiguoinnissa voisi tehota jopa täydellisiin homografeihin. Esimerkiksi verbi *lakata* on täydellinen homografi. Verbin paikkaluku kuitenkin riippuu sen merkityksestä. Jos kyse on toiminnan loppumisesta, verbi on yksipaikkainen. Jos taas kyse on suojaavan maalinesteen levittämisestä, verbi on kaksipaikkainen. Jos siis verbi ei saa lauseessa objektia, on kyse luultavimmin ensimmäisestä merkityksestä:

"Taistelut jatkuivat *lakkaamatta*."

Sataprosenttiseen disambiguointiin tuskin kuitenkaan päästäisiin tälläkään menetelmällä, sillä kaikki verbin vaatimat lauseenjäsenet eivät esiinny joka lauseessa, vaan ne saatetaan korvata tyhjillä ellipseillä.

Kaikenkaikkiaan vaikutti siltä, että suuri enemmistö homografeista olisi disambiguoitavissa yksinkertaisenkin syntaktisen analyysin avulla. Tutkimus paljasti myös, että ylivoimaisesti paras disambiguoija on ihminen. Yhtään sellaista homografia, jolle ihminen ei olisi pystynyt antamaan oikeata merkitystä, ei tässä tutkimuksessa kohdattu. Kun siis kehitetään tietokoneelle disambiguointiohjelmaa, myös ohjelmaa käyttävän ihmisen kannattaa antaa sanoa sanansa ongelmallisen homografin todellisesta merkityksestä.

Homografien disambiguoinnin toteuttaminen

Disambiguoinnin tarpeellisuus

Edellisissä luvuissa esiteltyjen tulosten perusteella homografit aiheuttavat suhteellisen vähän virheitä tekstihaussa. Ongelmaa voisi luonnehtia pikemminkin kiusalliseksi kuin haitalliseksi. Jos tekstihaun tarkkuutta haluttaisiin parantaa, eräiden muiden virhetyyppien eliminointi auttaisi huomattavasti enemmän kuin homografien disambiguointi.

Homografivirheet eroavat kuitenkin useimmista muista virhetyypeistä siten, että ne ovat suhteellisen helposti tunnistettavissa ja eroteltavissa omaksi ryhmäkseen. Monen muun virhetyypin väliset rajat on paljon vaikeampi määritellä. Vaikuttaa myös siltä, että homografien aiheuttamat virheet ovat nykyisenkin tietotekniikan korjattavissa, mitä ei voi sanoa useimmista tekstihaun ongelmista.

Disambiguoinnin toteutus tiedonhakujärjestelmässä

Homografien disambiguoinnin toteuttamiseksi lienee useitakin vaihtoehtoja, mutta yksinkertaisin tuntuisi olevan juoksevan tekstin lauseanalyysiin perustuva operaatio. Operaation toiminnalla on kuitenkin tiettyjä vaatimuksia, joiden pitää täyttyä ennen kuin operaatio voi toimia tyydyttävästi. Aivan kaikkia näistä vaatimuksista ei ehkä tarvitse täyttää, mutta mitä enemmän niistä toteutetaan, sitä varmemmin disambiguointi toimii oikein.

Ensimmäinen vaatimus on, että disambiguointiohjelman on tunnettava jokaisen homografin kaikki merkitykset. Ohjelman sanastossa on oltava myös homografisia erisnimiä. Jos ohjelma tuntee vain toisen homografin kahdesta perusmuodosta ja jos homografi esiintyykin tekstissä siinä toisessa merkityksessä, ei ohjelmalla ole valinnanvaraa: se tulkitsee homografin väärin. Tulisi olla mahdollista lisätä sanastoon myös uusia sanoja.

Toiseksi ohjelman on kyettävä tekemään syntaktinen analyysi jokaisesta tekstin lauseesta. Tämän analyysin on kuitenkin oltava joustava. Luonnollisen kielen teksti sisältää usein lauseita, jotka eivät ole kieliopin mukaisia. Tällaisia ovat sanomalehtitekstissä erityisesti otsikot. Otsikoita ei voi jättää tekstihaun ulkopuolellekaan, koska ne sisältävät usein keskeistä sanastoa. Ohjelman on siis kyettävä hyväksymään, että tekstissä saattaa olla sellaisiakin ilmaisuja, joita ei pysty jäsentämään lauseopin sääntöjen mukaan.

Tässä tutkimuksessa ilmeni, että tiedonhaun homografivirheissä on hyvin usein kyse erisnimen ja yleisnimen välisestä homografiasta. Näiden sanojen aiheuttamat hakuvirheet olisivat usein vältettävissä, jos ne pystyttäisiin tunnistamaan erisnimeksi ison alkukirjaimen perusteella. Uusimmissa tiedonhakujärjestelmissä tämä on jo mahdollista, joten osa homografivirheistä karsiutuu sen ansiosta. Virkkeiden alussa oleviin sanoihin tämä ei tosin tehoa, mutta jos sama sana esiintyy tekstissä isolla alkukirjaimella myös lauseen keskellä, se on todennäköisesti erisnimi. Alkukirjaimen lisäksi voisi ottaa huomioon, että erisnimi ei normaalisti esiinny monikossa, joten kyseisten sanojen monikkomuodot virkkeiden alussa voidaan tulkita yleisnimiksi.

Verbien paikkaisuuden määrittely olisi myös yksi homografiaan tehoavista toimenpiteistä. Jos disambiguointiohjelma kykenisi

tunnistamaan, onko verbi nolla-, yksi-, kaksi- vai kolmipaikkainen, se voisi karsia verbien homografioita. Jos homografisen verbin ympärillä on tietyt lauseenjäsenet, se voidaan disambiguoida sen paikkaluvun perusteella.

Jos tämän lisäksi ohjelma voisi vielä tunnistaa eräitä fraaseja, jotka koostuvat verbeistä ja adverbeista, olisi ongelmatapauksia vielä vähemmän.

Kaikkein parhaaseen disambiguointiin päästään, jos myös tietokonetta käyttävän ihmisen kyvyt otetaan käyttöön. Ihminen on ylivoimaisesti paras disambiguoija. Sellaisia homografeja, joiden merkitystä ihminen ei pystyisi tunnistamaan, on normaalissa tekstissä äärimmäisen vähän.

Uudemmissa tiedonhakujärjestelmissä käänteistiedosto muodostetaan usein siten, että dokumenttien sanat palautetaan perusmuotoon. Tämä tapahtuu suodattamalla tietokannan tekstit ensin perusmuotoihin palauttavan ohjelman läpi. Useimmat näistä ohjelmista eivät toistaiseksi pysty disambiguointiin, joten homografeille kirjataan käänteistiedostoon niin monta perusmuotoa kuin vaihtoehtoja on.

Tämän perusmuotoon palauttavan ohjelman tilalla voisi olla disambiguointiohjelma. Tämän ohjelman pitäisi sekä palauttaa sanat perusmuotoihin että disambiguoida homografit. Koska kyseessä on kuvitteellinen ohjelma, voimme olettaa, että siinä on täytetty kaikki esitellyt vaatimukset. Olisi hyödyllistä lisätä ohjelmaan myös oikolukumahdollisuus, niin kiusallisista kirjoitusvirheistä päästäisiin ainakin osittain.

Ohjelma voi toimia kahdella eri tavalla riippuen siitä, halutaanko disambiguoinnissa käyttää apuna ihmistyövoimaa vai ei. Jos disambiguoinnista halutaan suoriutua pelkästään konevoiman avulla, ohjelman toimintajärjestys voisi olla tämä:

1) Ohjelmalle syötetään tekstidokumentti, joka on tarkoitus lisätä tekstikantaan.

2) Ohjelma sijoittaa tekstikannan käänteistiedostoon kaikki dokumentin yksiselitteiset sanat.

Samalla ohjelma voisi oikolukea dokumentin ja pyytää käyttäjää hyväksymään tai korjaamaan tuntemattomat sanat. Jos sana on oikein kirjoitettu mutta outo oikolukuohjelmalle, ohjelma voi tallettaa sen muistiin ja lisäksi pyytää käyttäjää kertomaan sanan sanaluokan. Tämän tiedon avulla se voisi päätellä sanan taivutusmuodot, eikä sen tarvitse pyytää sanalle vahvistusta enää toistamiseen. (Tämän vaiheen voi jättää pois, mikäli käyttäjä ei halua tai ehdi puuttua ohjelman toimintaan.)

Tämän jälkeen jäljelle ovat jääneet vain homografit.

3) Ohjelma tekee dokumentista syntaktisen analyysin ja sijoittaa käänteistiedostoon kaikki analyysin perusteella disambiguoidut homografit. Disambiguoinnissa käytetään hyväksi erisnimien tunnistusta ja verbien paikkaisuuden määrittelyä. Myös adverbifraasien tunnistaminen tapahtuu tässä yhteydessä. Kun ohjelma huomaa homografin, joka on tulkittavissa sekä tietyksi adverbiksi että substantiiviksi, se tarkistaa, onko samassa lauseessa verbiä, jonka yhteyteen adverbi kuuluu. Jos on, se tulkitaan adverbiksi. Suurin osa homografeista disambiguoituu tässä vaiheessa.

4) Jäljellejääneitä homografeja tarkastellaan kutakin erikseen.

Jos tarkasteltava homografi on tulkittavissa kuuluvaksi kahteen tai useampaan eri sanaluokkaan, joista yksi on substantiivi, se tulkitaan substantiiviksi. Näin tehdään, koska substantiivi on tärkein sanaluokka tiedon haussa ja paremman saannin varmistamiseksi kyseenalaiset tapaukset on varminta tulkita substantiiveiksi. Jos tarkasteltava homografi on tulkittavissa kuuluvaksi kahteen tai useampaan eri sanaluokkaan, joista yksikään ei ole substantiivi ja joista yksi on verbi, se tulkitaan verbiksi. Verbit ovat toiseksi tärkein sanaluokka tiedonhaun kannalta. Myös muut sanaluokat voidaan panna "arvojärjestykseen". Hyvä järjestys olisi ehkä tämä: adjektiivi, numeraali, adverbi, postpositio-prepositio, interjektio, pronomini, konjunktio.

Jos homografi voidaan palauttaa kahteen tai useampaan samaan sanaluokkaan kuuluvaan perusmuotoon, on kaksi vaihtoehtoa: joko valitaan yleisempi sana tai sijoitetaan käänteistiedostoon molemmat sanat. Hyvän saannin varmistamiseksi olisi parempi sijoittaa käänteistiedostoon tunnistamattoman homografin kaikki perusmuodot.

5) Kun dokumentin kaikki sanat ovat käänteistiedostossa, ohjelman suoritus päättyy tai siirtyy seuraavaan dokumenttiin.

Ohjelman ei välttämättä tarvitse toimia juuri näin. Olisi ehkä parempi, jos vaiheet 2–4 voisi yhdistää samaan operaatioon, niin dokumenttia ei tarvitsisi käydä läpi useaan kertaan.

Jos ohjelmassa haluttaisiin hyödyntää ihmisen disambiguointikykyä, toimisi ohjelma osin eri tavalla. Vaiheet 1–3 sekä vaihe 5 eivät muuttuisi, mutta vaihe 4 menisi näin:

4) Kun ohjelma kohtaa homografin, jota se ei pysty disambiguoimaan, se tiedustelee neuvoa ohjelman käyttäjältä. Se antaa vaihtoehdot ja näyttää käyttäjälle lauseen, jossa homografi on, sekä mielellään muutaman ympäröivänkin lauseen. Käyttäjä päättelee tekstiyhteydestä homografin perusmuodon ja kertoo sen ohjelmalle.

Jos homografi on niin ongelmallinen, että edes ihminen ei pysty varmuudella disambiguoimaan sitä, hän voi joko antaa homografin kaikkien mahdollisten perusmuotojen mennä käänteistiedostoon tai arvata todennäköisimmän. Jos homografin merkitystä ei pysty tunnistamaan, on todennäköistä, että sillä ei ole juuri arvoa tiedon haussakaan.

Tämän toisen vaihtoehdon etuna on, että siten voitaisiin olla varmempia disambiguoinnin onnistumisesta. Huonona puolena on, että ohjelma olisi hitaampi ja raskaampi käyttäjälle, sillä hän joutuisi luultavasti tulkitsemaan homografeja melkein jokaisen dokumentin kohdalla. Paras olisi, että käyttäjälle olisi tarjolla samassa ohjelmassa molemmat vaihtoehdot, jotta hän voisi valita sopivan tavan tilanteen mukaan.

Kumpikaan näistä vaihtoehdoista tuskin olisi täysin virheetön. Luonnollinen kieli on liian monipuolista ja vaihtelevaa, jotta kukaan ihminen pystyisi hallitsemaan sitä täydellisesti – koneesta puhumattakaan. Jos kuitenkin siedetään, että virheitäkin voi joskus sattua, tässä luonnostellut ohjelmat parantaisivat nykyistä tilannetta.

Tiedonhakujärjestelmään olisi hyvä lisätä vielä mahdollisuus haluttaessa määrätä hakusanan sanaluokka. Jos esimerkiksi tiedonhakija hakee tietoja kuusista, hän voisi määrätä, että hakusanan on haettava vain substantiiveja eikä *kuusi*-numeraaleja. Toisaalta, jos tiedonhakija hakee tietoa esimerkiksi nuorista tai venäläisistä, hänelle on luultavasti yhdentekevää, esiintyvätkö käsitteitä vastaavat hakusanat tekstissä substantiiveina vai adjektiiveina.

Disambiguoinnin vaikutukset

Jos edellä kuvattu disambiguointiohjelma olisi käytössä tiedonhakujärjestelmässä, sen välitön seuraus olisi, että osahomografiongelma katoaisi lähes kokonaan. Aivan kokonaan homografeista tuskin kuitenkaan päästäisiin, sillä virheitä voi aina sattua. Sitä paitsi se ei auttaisi täydellisiin homografeihin, paitsi erisnimiin, jos ne tunnistettaisiin erisnimeksi ison alkukirjaimen perusteella, sekä joihinkin verbeihin, jos ne pystytään disambiguoimaan paikkaisuuden perusteella.

Ohjelman negatiivisia vaikutuksia taas olisi lähinnä se, että uusien dokumenttien lisääminen tietokantaan olisi työläämpää. Dokumentin analysointi voi kestää pitkään. Vielä enemmän aikaa menee, jos ohjelma pyytää käyttäjää disambiguoimaan homografeja. Tämä vaatii käyttäjältä aktiivista ohjelman toiminnan seuraamista sekä myös kielen ja kieliopin tuntemusta.

Itse tiedonhaun tuloksiin ohjelmalla olisi vaihteleva vaikutus. Suurimpaan osaan hauista se ei vaikuttaisi lainkaan, mutta joissakin hauissa tarkkuus voisi parantua merkittävästikin. Saannin ei pitäisi heikentyä. Tosin on teoriassa mahdollista, että homografinen hakusana esiintyy relevantissa dokumentissa vain virheellisessä merkityksessään ja dokumentti jää siksi löytymättä disambiguoinnin jälkeen. Tällainen tapaus ei liene kuitenkaan kovin todennäköinen.

Disambiguointiohjelmalla varustetussa tiedonhakujärjestelmässä ei tulosjoukkoihin tulisi myöskään katkaisuvirheitä. Tämä on perusmuotoisen käänteistiedoston etu joka tosin voidaan toteuttaa ilman homografien disambiguointiakin.

Yhteenveto

Tämän tutkimuksen tavoitteina oli sekä kartoittaa homografiongelman yleisyyttä

tekstihaussa että pohtia keinoja ongelman ratkaisuksi. Tekstihaun homografiongelmaa ei ollut aiemmin tutkittu, mutta nyt tiedetään enemmän ongelman laajuudesta ja merkittävyydestä sekä sen ratkeavuudesta suomenkielisissä tekstikannoissa.

Homografia osoittautui suhteellisen vähäiseksi ongelmaksi tekstihaussa, kuten edeltävätkin tutkimukset ennakoivat. Homografien disambiguointi parantaisi hakujen tarkkuutta vain hivenen jos ollenkaan. Ongelma ei ole sitä suuruusluokkaa, että sen ratkaisemiseksi kannattaisi uhrata resursseja merkittävästi.

Lauseenjäsennysohjelmia on kuitenkin tekeillä joka tapauksessa, sillä niitä tarvitaan paljon muuallakin kuin tiedonhaussa. Tekoälyn tutkimuksen tavoitteisiin kuuluu saada tietokone ymmärtämään ihmisen kieltä, ja automaattisen lauseenjäsennyksen toteuttaminen on tässä tärkeä välitavoite. Tietokonelingvistiikan kehittämisen myötä voi hyvinkin syntyä ohjelmia, joita voitaisiin hyödyntää myös tiedonhaussa.

Homografien disambiguointi ei ehkä vaatisi kovin ihmeellisiä operaatioita. Melkein kaikki homografit disambiguoituisivat, mikäli tiedonhakujärjestelmään yhdistettäisiin toimiva lauseenjäsennysohjelma. Tämä olisi varmasti nykypäivänkin tietotekniikan toteutettavissa.

Tämä tutkimus on osoittanut, että homografia on vähäinen ongelma tekstihaussa. Se on kuitenkin myös osoittanut, että se on korjattavissa oleva ongelma ja että sen korjaamisesta olisi lähinnä positiivisia seurauksia. Vaikka esitelty disambiguointimenetelmä ei ehkä ratkaisisikaan ongelmaa sataprosenttisesti, se voisi toimia ainakin väliaikaisratkaisuna, kunnes tietokonelingvistiikka on kehittynyt pitemmälle. Homografien aiheuttamat ongelmat on pakko ratkaista tavalla tai toisella, jos tietokone halutaan ikinä saada ymmärtämään luonnollista kieltä.

Hyväksytty julkaistavaksi 2.7.1996.

Lähteet

Alkula, R. & Honkela, T. 1992. Tekstin tallennus- ja hakumenetelmien kehittäminen suomen kielen tulkintaohjelmien avulla. FULLTEXT-projektin loppuraportti. VTT. Espoo.

Karlsson, F. 1994. Yleinen kielitiede. Gaudeamus. Helsinki.

Kristensen, J. & Järvelin, K. 1990. The effectiveness of a searching thesaurus in freetext searching in a full-text database. Int. Classif., 17 (2). p. 77–84.

Laalo, K. 1990. Säkeistä patoihin: Suomen kielen monitulkintaiset sanamuodot. Suomalaisen kirjallisuuden seura. Vaasa.

Leppänen, E. 1995. Osahomografien disambiguoinnin vaikutukset ja toteuttaminen tekstihaussa. Informaatiotutkimuksen laitos. Tampereen yliopisto. Pro gradu -tutkielma.

Saukkonen, P., Haipus, M., Niemikorpi, A. & Sulkala, H. 1982. Suomen kielen homonyymeja. Språkhistoria och språkkontakt i Finland och Nord-Skandinavien. Kungl. skytteanska samfundets handlingar. Nr 26. s. 255–272.

Sormunen, E. & Alkula R. 1990. Suomenkielisten tekstitietokantojen tallennus- ja hakutekniikkojen kehittäminen. Esitutkimusraportti. VTT. Espoo.

Sormunen, E. 1994. Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanomalehtiaineistoa sisältävässä tekstikannassa. Informaatiotutkimuksen laitos. Tampereen yliopisto. Lisensiaattityö.