Ari Pirkola

Anaforiset ilmaisut kielitieteen, tietokonelingvistiikan ja informaatiotutkimuksenkin ongelmana

Pirkola, Ari, Anaforiset ilmaisut kielitieteen, tietokonelingvistiikan ja informaatiotutkimuksenkin ongelmana [Anaphora as a research problem in linguistics, computer linguistics and information studies]. Kirjastotiede ja informatiikka 13 (4): 120–132, 1994.

The article reviews literature on anaphora in the fields of linguistics, computer linguistics and information science. Anaphora are textual elements (most often pronouns) which refer to earlier text elements (called correlates) and share the meaning of the correlates. This traditional and established definition as a basis the concept of anaphora is discussed in detail. The emphasis is on anaphora resolution – determining the textual element to which an anaphoric expression refers. Resolution methods based on syntactic, semantic and discourse level text processing are presented. Included are also studies on the frequency of anaphora in the text and the linear distance between the pronoun and its correlate. Anaphora as a problem of IR research is also discussed. From the literature reviewed two conclusions are drawn. Firstly, it is not likely that a comprehensive resolution system for unrestricted natural language text could be built in the near future. However, restricted types of resolution systems should be considered, because anaphoric resolution may achieve a high degree of correctness in some text types, anaphora classes and correlate types. Secondly, because natural language is essential in the process of information retrieval, IR research should be in close association with linguistic research.

Address: Takalankuja 2 E 37, FIN-40740 Jyväskylä, Finland.

1. Ongelma ja sen avaaminen

Kun pari vuotta sitten me tulevat informaatiotieteilijät kävimme Kalervo Järvelinin tekstitietokannat ja niiden hakumenetelmät -kurssin, opimme mm. sen, että monesti kokotekstihaussa kannattaa ennemmin käyttää läheisyysoperaatiota kuin "ANDiä". Pitkissä dokumenteissa läheisyysoperaatiolla taataan hakuavaimien kuuluminen samaan asiayhteyteen paremmin kuin leikkausoperaatiolla. Siis kun haluan tietää, mitä arvioita Sixten Korkman on tehnyt taloudellisesta kehityksestä, niin relevantteja dokumentteja löydän paremmin lausekkeella paragraph ((sixten korkman*), arvio*) kuin lausekkeella (sixten korkman*) AND arvio*. Mutta entä jos kirjoittaja on koko nimen tilalla käyttänyt pronominia, ilmaisten asian esim. sanoilla *hän arvioi,* niin että kirjoituksessa koko nimi ja toinen hakuavain eivät kertaakaan ole samassa kappaleessa. Eihän paragraph-operaatiosta silloin hyötyä ole, vaan asiahan on päinvastoin. Mitä se Järvelin oikein puhui? No, tottahan hän tietenkin puhui, ja ristiriidassa onkin kyse siitä, että

kun ollaan tekemisissä luonnollisen kielen kanssa, niin hyvässäkin ja tilanteeseen sopivassa menetelmässä – läheisyysoperaatiohaku ja hakutuloksen tarkkuus – on aina puutteensa. Konkreettisesti esimerkissä on kyse siitä, että hakuavaimen paikalla on pronomini, anafora. Kyse on myös siitä, että tiedonhakijalle anaforat ovat rikka kielen rikkauden rokassa, rikkauden joka sinänsä on arvokasta, mutta tietoa tarvitsevalle usein tuskastuttava asia. Kyse on siitä, että anaforilla on käytännön merkitystä kokotekstihaussa. Kyse on myös siitä, ne tulee ottaa huomioon niin tiedonhaussa kuin tiedonhaun tutkimuksessakin.

Anafora on tekstin elementti, joka viittaa aikaisempaan tekstinkohtaan ja joka on tämän kohdan kanssa samaviitteinen eli samatarkoitteinen (Halliday ja Hasan 1976). Suurin osa tekstin anaforista on pronomineja, mutta myös muihin sanaluokkiin kuuluvilla sanoilla voidaan viitata anaforisesti. Anaforaa tutkitaan niin teoreettisessa kielitieteessä kuin tietokonelingvistiikassakin. Ratkaisua on etsitty varsinkin kysymykseen, minkälaisten rajoitteiden alainen anaforan ja sen korrelaatin suhde on. Korrelaatti on se sana, sanaliitto tai tekstijakso, johon anafora viittaa. Tutkijoita kiinnostaa myös se, millä tavalla avaaminen eli resoluutio, ts. anaforan korvaaminen korrelaatillaan, voitaisiin suorittaa automaattisesti. Avaaminen on tarpeellista, jotta tiedettäisiin, mitä anafora kussakin tilanteessa "merkitsee". Käytännöllisenä tavoitteena on nykyistä suorituskykyisimpien luonnollista kieltä käsittelevien (natural language processing, NLP) järjestelmien kehittäminen esim. täysin automaattiseen koneelliseen kielenkääntämiseen ja tiedonhakuun.

Edellä käytännön esimerkin avulla perusteltiin anaforan tutkimuksen tarpeellisuutta tiedonhaun alueella. Asiaa voidaan ja tuleekin tarkastella myös toisesta, yleisemmästä lähtökohdasta käsin. Voidaan katsoa, mitkä ovat tiedonhaun tutkimuksen tavoitteet. "Tiedon tallennuksen ja haun tutkimuksen perimmäinen tavoite on kehittää käsitteitä, menetelmiä ja järjestelmiä, joiden avulla kaikki tieto, olipa se missä tahansa muodossa ja missä tahansa paikassa, saadaan vaivattomasti kenen tahansa sitä tarvitsevan ulottuville ja siten esitettynä, että tarvitsijan on mahdollisimman helppo tämä tieto omaksua." "Tiedonhaku luonnollisella kielellä siten, että löydetään kaikki tarvittu tieto tarvitsematta vastaanottaa suurta määrää irrelevanttia tietoa, aiheuttaa suuria vaikeuksia." (Järvelin 1992.) Koska luonnollinen kieli kuuluu olennaisena osana tiedonhakuun, sitähän käytetään sekä hakijan kysymyksissä että informaation esittämisessä, ovat tiedonhaun ongelmat pitkälle kielellisiä. Pyrittäessä tiedonhaun tutkimuksen perimmäiseen tavoitteeseen on lingvistisen tutkimuksen liityttävä kiinteästi tiedonhaun tutkimukseen. Anafora on yksi lingvistisistä tekijöistä, jonka vaikutusta tiedonhakuun on pyrittävä selvittämään, sillä tiedonhaun alueella anaforilla on merkitystä mm. järjestelmissä, joissa tekstin semanttisella esittämisellä on keskeinen sija, sekä hakujärjestelmissä, joissa lasketaan sanojen painoarvot (Liddy, Bonzi, Katzer & Oddy 1987). Anaforilla sekä ellipseillä eli vaillinaisilla ilmauksilla on vaikutusta myös läheisyysoperaatiohakujen tulokseen. Pro gradu -työni (Pirkola 1994) koski tätä aihetta.

Esittelen tässä kirjoituksessa työni anaforia koskevan kirjallisuusosan. Tämän kielitieteellisen ja tietokonelingvistisen kirjallisuuden avulla pyrin tuomaan esille sen, mitä käsitteellä anafora tarkoitetaan sekä millä menetelmillä sen avaaminen voidaan tehdä. Tarkastelen myös tiedonhaun anaforaa tutkivaa kirjallisuutta.

2. Anaforan määrittely

Edellä tuli jo esille, miten anafora perinteisesti määritellään. Tässä perinteisessä ja vakiintuneesa määritelmässä otetaan huomioon sekä tekstin sisäiset suhteet että ulkomaailma. Hirst (1981) niin ikään selittää anaforan ja sen korrelaatin samaviitteisiksi, mutta hän lisää, että anafora sisältää vähemmän informaatiota kuin sen korrelaatti.

Hakulisen ja Karlssonin (1988) mukaan käsitteeseen anafora sisältyy paitsi samatarkoitteisuus myös samamerkityksisyys. He esittävät seuraavan esimerkin.

1. Kallella ei ole koiraa, mutta Pekalla on.
   Se on bokseri.

Ellipsi on korvannut edellisen lauseen NP:n *koiraa* kanssa samamerkityksisen NP:n, johon pronomini tuntuisi viittaavan. Pintarakenteen kannalta kuitenkin ainoa korrelaatti on pronominin kanssa eritarkoitteinen. (Hakulinen ja Karlsson 1988.) Tällaista Hakulinen ja Karlsson (1988) kutsuvat *merkitysanaforaksi,* ja siinä on kyse intension samuudesta taikka ainakin lähisukulaisuudesta. Hirst (1981) erottaa kaksi anaforatyyppiä: IRA (identity of reference anaphora) ja ISA (identity of sense anaphora). IRAlla hän tarkoittaa sitä, että anafora ja sen korrelaatti viittaavat juuri samaan

olioon. ISAssa puolestaan on kyse esimerkin 1 kaltaisista tilanteista, joissa anafora ja sen korrelaatti Hirstin mukaan ovat samamerkityksisiä mutta eivät samatarkoitteisia. Itse asiassa esimerkissä 1 on kyse lähimerkityksisyydestä, ja oikeampaa olisikin puhua toisaalta samamerkityksisyydestä ja toisaalta lähimerkityksisyydestä. Edellisten lisäksi käsitteeseen anafora sisältyy *tämä*-pronominin osoitteleva käyttö, mitä kutsutaan *tekstuaaliseksi deiksikseksi* (Hakulinen ja Karlsson 1988). Tästä esimerkkinä ovat seuraavat virkkeet.

2. Pienille valittakoon hiukan yksinkertaisempaa runoutta.
Tähän tarkoitukseen käy esim. Aroseniuksen kuvakirja Kissamatka.
(Hakulinen ja Karlsson 1988)

Yleensä samaviitteisyydellä tarkoitetaan samatarkoitteisuutta. Koska anaforassa ei aina ole kyse samasta tarkoitteesta, Hakulinen ja Karlsson (1988) käyttävät anaforasta vain termiä *samaviitteinen*, jonka he kuvaavat *samatarkoitteista* väljemmäksi.

Pronominien lisäksi myös substantiivit voivat toimia anaforina. Tällöin on kyse geneerisistä sanoista, jotka ovat merkitykseltään laajempia kuin korrelaattinsa ja täten sitä ylempänä hierarkiassa. Esimerkiksi sanaa *maa* käytetään usein, kun viitataan tekstissä edellä olevaan valtion nimeen. Tämä geneerinen sana voidaan aina vaihtaa pronominiin. Nämä eivät kuitenkaan ole aivan identtisiä, sillä substantiivisen anaforan avulla kirjoittajalla on mahdollisuus tuoda esille tietty näkökulma aiheeseen, varsinkin kun substantiivin eteen voidaan aina laittaa määrite. (Halliday ja Hasan 1976.)

Hirstin (1981) mukaan ei ole selvää, missä anafora loppuu, ja esim. parafraasin ja anaforan raja on hänen mukaansa diffuusi. Lähellä anaforaa on myös substituution käsite. Sillä tarkoitetaan tekstin joidenkin lausekkeiden tai jaksojen korvaamista. Hakulisen ja Karlssonin (1988) mukaan siinä ei ole kyse niinkään samatarkoitteisuudesta eikä varsinaisesta intension samuudestakaan kuin tietyn leksikaalisen aineksen korvaamista toisella, substituutilla. Nominaalisen aineksen yleisin substituutti on *sellainen*. Sanaa *sitä* käytetään myös usein substituuttina. (Hakulinen ja Karlsson 1988.) Virke 3 on esimerkkinä substituutiosta.

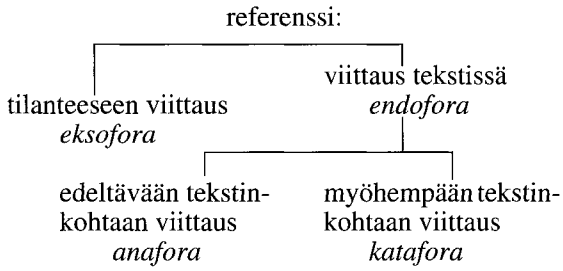
3. Minä olen ikuinen kakkonen, *jos sitäkään*.
(Hakulinen ja Karlsson 1988)

Uudempaa näkemystä anaforasta edustavat Carter (1987) ja van Deemter (1992). He lähtevät siitä, että anafora on tekstin sidoskeino. Hyvin muodostuneessa tekstissä lauseiden täytyy jotenkin liittyä eli olla sidoksissa toisiinsa, ja anafora on yksi sidoskeinoista (Halliday ja Hasan 1976; Karlsson 1982). Carterin (1987) mukaan anafora on merkitykseltään tyhjä tai epätäydellinen, jos sitä tarkastellaan erillään muista lauseen elementeistä, ja se voidaan ymmärtää vain tarkastelemalla myös sitä tekstin elementtiä, jonka kanssa sillä on sidoksisuussuhde. Van Deemterin (1992) mukaan perinteinen käsitys anaforasta sisältyy Carterin määritelmään. Sekä Carterin että van Deemterin tarkoituksena on laajentaa anaforan käsite koskemaan myös sellaisia tapauksia, joissa ei ole kyse samaviitteisyydestä. Heidän mukaansa muut kuin samaviitteiset tekstiä sitovat elementit voivat olla anaforia, niin kuin *the author* seuraavassa esimerkissä. Esimerkin viittaussuhteen ymmärtäminen edellyttää tietoa siitä, että kirjoilla on kirjoittajansa. On huomattava, ettei tämä uudempi näkemys anaforasta ole syrjäyttänyt perinteistä käsitystä.

4. John read a *book about Schubert.*
He wrote a letter to *the author.*
(van Deemter 1992)

3. Anaforan luokittelu

Halliday ja Hasan (1976) luokittelevat viittaussuhteet seuraavasti:



Eksoforat viittaavat tekstin ulkopuolella olevaan tilanteeseen, ja endoforat ovat tekstin sisäisiä viittauksia. Endoforista katafora edeltää korrelaattia ja anafora tulee sen jälkeen. Toisinaan anaforasta puhutaan laajasti, siten että myös katafora katsotaan käsitteeseen kuuluvaksi.

Fraurud (1988) erottaa toisistaan tilanteet, joissa anafora viittaa nominaalilausekkeeseen ja tilanteet, joissa se viittaa laajempaan tekstijaksoon. Tämä voi olla virkettäkin laajempi jakso. ISA- ja IRA-jaottelun lisäksi Hirst (1981) jakaa anaforat sen mukaan, esiintyvätkö korrelaatit eksplisiittisesti

tekstissä vai eivät. Jälkimmäisestä tilanteesta on kyse seuraavassa esimerkissä.

5. Ross gave each girl a crayon. *They* used *them* to draw pictures of Daryel in the bath. (Hirst 1981)

Pronominin *they* viittauksen kohde on *girls* ja pronominin *them crayons*. Kumpikaan näistä ei kuitenkaan ole tekstissä, vaan kohteiden *girls* ja *crayons* olemassaolo on pääteltävä annetuista elementeistä *each girl* ja *crayon*.

Liddyn ym. (1987) kehittämä luokittelu on esitetty alla. He keräsivät yhteen kaikki mahdolliset sanat, jotka englannin kielessä voivat toimia anaforina. Kirjallisuudesta he eivät löytäneet minkäänlaista yhtenäistä luokittelua, joten seuraavassa esitettävä on tiettävästi ensimmäinen kattava anafora-luokittelu.

1. Pronominit (Central pronouns)
   a. Persoonapronominit
   (Personal pronouns – he, him, she, her, it)
   b. Possessiivipronominit
   (Possessive pronouns – his, her, their)
   c. Refleksiivipronominit
   (Reflexive pronouns – itself, themselves)
2. Demonstratiivipronominit
   (Nominal demonstratives – this, these, those)
3. Relatiivipronominit
   (Relative pronouns – who, which, what)
4. Nominin substituutit
   (Nominal substitutes – above, former, one)
5. Apuverbi (Proverb – do)
6. Indefiniittipronominit
   (Indefinite pronouns – any, each, many)
7. Adjektiivit (Proadjectives – another, identical)
8. Adverbiaalit (Proadverbials – so, such, similarly)
9. Substantiivit (Subject references – S, Ss)
10. Määrätty artikkeli (Definite article – the)

Valitettavasti kirjallisudessa ei ole ollut arvioita tästä luo kittelusta (Liddy 1990). Liddyn ym. esittämien luokkien nimet on edellä suomennettu, mutta on huomattava, että kaikissa tapauksissa suomen kielen luokka ei ole aivan identtinen vastaavan englannin kielen luokan kanssa. *Subject references* -termillä Liddy ym. tarkoittanevat substantiiveja, mutta artikkelista ei valitettavasti käy ilmi, mitä käsitteellä tarkkaan ottaen tarkoitetaan.

4. Anafora tekstissä

Taajuus

Liddy ym. (1987) laskivat anaforien määrän kahden tietokannan, PsycINFO:n ja INSPEC:in, abstrakteista. Edellinen käsittelee käyttäytymistieteitä ja jälkimmäinen tietojenkäsittelytiedettä ja tekniikkaa. PsycINFOssa anaforia oli keskimäärin 4.49 ja INSPECissä 2.86 abstraktia kohden. Keskiarvoksi näistä saadaan 3.67.

Fraurud (1988) selvitti anaforisten pronominien jakaumaa ruotsin kielessä korrelaatin tyypin mukaan. Hän erotti korrelaatit, jotka ilmaisevat henkilöä, ne jotka ilmaisevat objektia sekä korrelaatit, jotka ovat laajempia tekstijaksoja kuin nominaalilauseke (propositionaalinen korrelaatti). Henkilöihin viitataan pääasiassa pronomineilla *han* ja *hon* ja muissa viittauksissa käytetään pääasiassa pronomineja *den* ja *det*. Aineisto koostui ruotsinkielisistä kertomuksista, raporteista ja teknisiä keksintöjä käsittelevistä artikkeleista. Näistä jokaisesta Fraurud analysoi 200 anaforista pronominia sisältävän tekstijakson. Tulokset on esitetty taulukossa 1.

Kertomuksissa ja raporteissa pronominit ovat hyvin usein henkilöpronomineja, artikkeleissa taas

		Kertomukset	Raportit	Artikkelit	Yhteensä
Henkilö	N	186	157	12	355
	%	93.0	78.5	6.0	59.2
Objekti	N	11	34	149	194
	%	5.5	17.0	74.5	32.3
Propositionaalinen	N	3	9	39	51
	%	1.5	4.5	19.5	8.5
Yhteensä	N	200	200	200	600
	%	100.0	100.0	100.0	100.0

Taulukko 1. Pronominien jakauma korrelaatin tyypin mukaan (Fraurud 1988)

objekteihin viittaavia.Fraurudin mukaan tämä on luonnollista, koska kertomuksissa ja raporteissa käsiteltiin ihmisiä ja artikkeleissa esineitä ja asioita. Hän pitää kuitenkin havaintoa mielenkiintoisena. Artikkeleissa nimittäin kaikkien pronominien määrä oli vähäinen, ja näin ollen pronominit *han* ja *hon,* mukaanlukien näiden objektimuodot *honom* ja *henne* sekä possessiivimuodot *hans* ja *hennes,* ovat kaiken kaikkiaan huomattavasti yleisempiä kuin pronominit *den* ja *det* ja niiden objekti- ja possessiivimuodot.

Lineaarinen etäisyys korrelaatista

Kun mahdollisia korrelaattikandidaatteja määritetään, on tiedettävä, kuinka etäällä korrelaatti voi sijaita, missä se todennäköisimmin sijaitsee ja mitkä eri seikat vaikuttavat sen sijaintiin (Fraurud 1988). Fraurud (1988) tutki pronominin ja sen korrelaatin lineaarista etäisyyttä aineistonaan samat tekstityypit kuin pronominien jakaumatutkimuksessaan. Raporteissa ja artikkeleissa pronomini ja sen korrelaatti olivat useammin samassa virkkeessä kuin peräkkäisissa virkkeissä. Päinvastainen on tilanne kertomuksissa. Ottaen huomioon kaikki kolme tekstityyppiä noin 90 %:a korrelaateista sijaitsee joko samassa tai edellisessä virkkeessä kuin pronomini. Vastaava luku Hobbsin (1978) arkeologista tekstiä, sanomalehtiartikkeleita ja romaanitekstiä käsittävässä englanninkielisessä aineistossa oli 98 %. Fraurudin (1988) mukaan kolme seikkaa näyttää lisäävän lineaarista etäisyyttä: henkilöpronominit, henkilö on tarinan keskeinen hahmo ja upotettu diskurssirakenne. Viimeksi mainitulla seikalla tarkoitetaan sitä, että pronominin ja sen korrelaatin välillä on tekstijakso, joka sisältää suoran lainauksen.

5. Anaforan avaaminen

Automaattisen avaamisen voidaan ajatella muodostuvan kahdesta vaiheesta: ensiksi haetaan kaikki korrelaattivaihtoehdot, joista sitten valitaan oikea tai todennäköisin korrelaatti (Fraurud1988). Tietokoneen kannalta avaamiseen liittyy usein monitulkintaisuuden ongelma: korrelaattivaihtoehtoja on enemmän kuin yksi. Ei ole kovinkaan harvinaista, että viittaussuhteet esitetään tekstissä epäselvästi. Seuraavassa esimerkissä ei tietokone sen paremmin kuin ihminenkään ilman lisäinformaatiota pystyisi löytämään oikeaa korrelaattia.

6. Ross told Daryel *he* had passed the exam.
(Hirst 1981)

Yleisempää on kuitenkin, että on olemassa "oikea" korrelaatti,mutta mikään selvä rajoite ei pysty sitä osoittamaan. Usein tällaisissa tilanteissa tarvitaan päättelyä, johon vaadittava informaatio voi olla runsasta ja liikkuu usealla kielen tasolla.

Avaaminen voidaan suorittaa erilaisista, kielen eri tasoihin liittyvistä lähtökohdista käsin. Anafora on morfologisen, syntaktisen, semanttisen ja pragmaattisen tason käsite (Akmajian,Demers, Farmer & Harnish 1990). Pelkästään yhteen lähtökohtaan perustuva menetelmä voi antaa kohtalaisen hyviä tuloksia (esim. Fraurud 1988). Järjestelmän, joka pyrkii 100 % avaamistulokseen, tulee kuitenkin, syntaktisten rakenteiden, diskurssin rakenteen ja sanojen merkityksen tunnistamisen lisäksi, pystyä päättelemään ja siinä tulee olla esitettynä yleistä maailmaan liittyvää tietoa (Hirst 1981). Tällaisen rajoittamattomasti kaikkea kirjoitettua tekstiä käsittelevän järjestelmän kehittäminen ei liene ainakaan lähitulevaisuudessa mahdollista. Rajatuilla aihealueilla hyvin toimivien luonnollista kieltä käsittelevien järjestelmien kehittäminen sitä vastoin on mahdollista (Grishman 1986).

Seuraavassa tarkastellaan viittä keskeistä lähestymistapaa ongelmaan. Tässä käsiteltävien lisäksi muitakin lähestymistapoja on (ks. Hirst 1981). Anaforan avaamiseen liittyvä tutkimus on ollut teoriapainotteista. Hobbs (1978) ja Fraurud (1988) ovat kuitenkin testanneet kehittämiensä algoritmien toimivuutta käytännössä. Syntaksin analyysin yhteydessä tarkastellaan heidän töitään.

Pronominityypin hyväksikäyttö avaamisessa

Avaamisessa voidaan käyttää hyväksi sitä, että henkilöön viitataan eri pronominilla kuin muihin olioihin. Samoin korrelaatin luku ja useissa kielissä suku sekä korrelaatin ilmaiseman henkilön sukupuoli määräävät, mitä pronominia käytetään. Tieto näistä ei kuitenkaan auta, jos kaksi tai useampia korrelaattivaihtoehtoja kuuluu samaan kategoriaan. (Fraurud 1988; Hobbs 1978.) Lisäksi

poikkeukset eivät ole kovinkaan harvinaisia. Englannin kielessä pronominilla *she* voidaan viitata esimerkiksi laivaan.

Syntaktinen analyysi: syntaktiset rajoitteet

Syntaktisen analyysin suorittaviin luonnollista kieltä käsitteleviin järjestelmiin kuuluu sanakirja, kielioppi ja jäsennin. Sanakirjassa kunkin sanan kohdalla on se sanaluokka tai muu kategoria, mihin se kuuluu. Jäsennin siihen liittyvine kielioppeineen analysoi lauseen ja muodostaa sen syntaktisen rakenteen kuvauksen. Lausekerakennekieliopit muodostavat konstituenteista (ks. sanasto) koostuvan hierarkkisen puurakenteen. Syntaktisiin rajoitteisiin perustuva anaforan avaaminen suoritetaan tässä puurakenteessa.

Noam Chomskyn tutkimuksista alkoi kielentutkimuksen perinne,jonka piirissä on kehitetty sääntöjä, jotka pyrkivät täsmällisesti esittämään, milloin pronomini on tai ei ole samaviitteinen lauseen jonkin toisen elementin kanssa. Anaforan avaamisessa tällaisten syntaktisten rajoitteiden avulla pyritään löytämään oikea korrelaatti tai karsimaan mahdollisten korrelaattien joukko mahdollisimman pieneksi.

Lees ja Klima (1963) esittivät, että lauseen (simplex sentence) subjekti ja objekti ovat samaviitteisiä vain, jos objekti on refleksiivipronominina.

7. The boys looked at them.
8. The boys looked at themselves.

(Lees ja Klima 1963)

Lauseessa 8 pronomini on samaviitteinen subjektin kanssa, mutta lauseessa 7 se on eriviitteinen. Fraurudin (1988) ja Hobbsin (1978) algoritmeihin sisältyi tämä rajoite (ks. s. 126–127).Jackendoffin (1972) mukaan seuraavat lauseet ovat ristiriidassa Leen ja Kliman kuvaaman säännön kanssa.

9. Tom told Dick Harry's story about himself.
10. Tom told Dick Harry's story about him.

Esimerkissä 9 *himself* ei viittaa nimeen *Tom* vaan nimeen *Harry.* Esimerkissä 10 *him* viittaa joko nimeen *Tom* tai *Dick.* Kirjallisuudessa on muitakin tämänkaltaisia esimerkkejä, joissa pronomini tai refleksiivipronomini preposition jäljessä viittaa muuhun tekstin elementtiin, kuin mihin sen Leen ja Kliman esittämän säännön mukaan tulisi viitata. On kuitenkin vaikeaa elle imahdotonta keksiä esimerkkiä siitä, että pronomini on varsinaisessa objektiasemassa samaviitteinen lauseen subjektin kanssa.

Langackerin (1969) mukaan pronomini ei voi sekä edeltää että määrätä (command) korrelaattiaan. Määräämissuhteen Langacker (1969) määrittelee seuraavasti: solmu A määrää solmua B, jos on voimassa, (1) etteivät A ja B dominoi toisiaan ja (2) että se S-solmu, joka välittömimmin dominoi A:ta, dominoi myös B:tä. Sääntö selittää pronominien viittaussuhteet seuraavissa virkkeissä.

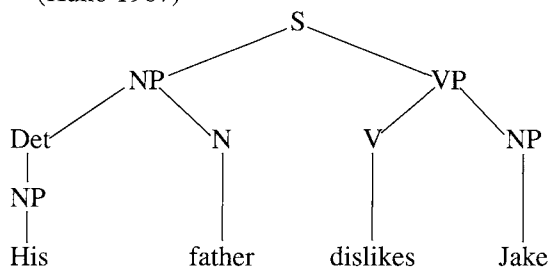
11. Jake left town after he robbed the bank.
12. He left town after Jake robbed the bank.
13. After Jake robbed the bank, he left town.
14. After he robbed the bank, Jake left the town.

(Jackendoff 1972; Grishman 1986)

Langackerin esittämän säännön mukaan *Jake* voi olla pronominin *he* korrelaatti virkkeissä 11, 13 ja 14, muttei virkkeessä 12. Jälkimmäisessä nimittäin pronomini sekä edeltää että määrää NP-solmua *Jake* (Grishman 1986). Säännön avulla voidaan joitakin korrelaattivaihtoehtoja sulkea pois, mutta se ei kuitenkaan kerro, mikä jäljelle jäävistä vaihtoehdoista on korrelaatti (Grishman 1986). Seuraavassa lauseessa *his* ja *Jake* ovat samaviitteisiä, ja pronomini *his* sekä edeltää että määrää solmua *Jake,* eikä sääntö näin ollen pidä paikkaansa.

15. His father dislikes Jake.

(Kuno 1987)

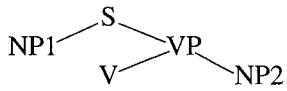


(Kuno 1987)

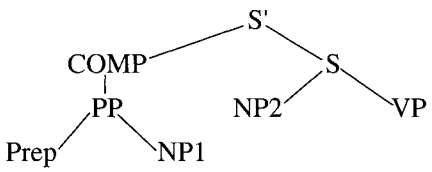
Myös Reinhart (1983) selittää samaviitteisyyttä määräämisellä.Hän käyttää termiä "c-command" (constituent-command), ja kyse on eri käsitteestä kuin edellä. Määräämisen Reinhart (1983) määrittelee seuraavasti: Solmu A määrää solmua B vain, jos ensimmäiseksi haarautuva solmu C, joka dominoi A:ta, joko (1) dominoi B:tä tai (2) sitä välittömästi dominoi solmu D, joka dominoi B:tä, ja D on samaa kategoriatyyppiä kuin C.

C-command-käsitteen havainnollistamiseksi Kuno (1987) esitti seuraavat puurakenteet.

a.



b.



Rakenteessa a NP1 määrää NP2:ta, koska S, joka on ensimmäiseksi haarautuva solmu, joka dominoi NP1:tä, dominoi NP2:ta.Sen sijaan NP2 ei määrää NP1:tä, koska VP, joka on ensimmäiseksi haarautuva solmu, joka dominoi NP2:ta, ei dominoi NP1:tä.

Rakenteessa b NP1 ei määrää NP2:ta, koska PP, joka on ensimmäiseksi haarautuva solmu, joka dominoi NP1:tä, ei dominoi NP2:ta. Sitä vastoin NP2 määrää NP1:tä, koska S, joka on ensimmäiseksi haarautuva solmu, joka dominoi NP2:ta, on välittömästi S':n dominoima, ja tämä puolestaan dominoi NP1:tä, ja S' kuuluu samaan kategoriatyyppin kuin S.

Reinhartin (1983) mukaan NP1 ja NP2 eivät voi olla samaviitteisiä, jos NP1 määrää NP2:ta ja NP2 ei ole pronomini. Katsotaan esim. seuraavia virkkeitä.

16. *Nicholas* left after *he* found the tricycle.
17. He left after Nicholas found the tricycle.
18. After *he* found the tricycle, *Nicholas* left.
(Akmajian ym. 1990)

Esimerkkivirkkeissä 16 ja 18 *he* ja *Nicholas* ovat samaviitteisiä, toisin kuin virkkeessä 17, jossa pronomini määrää solmua *Nicholas* (Akmajian ym. 1990).

Akmajian ym. (1990) pitävät Reinhartin esittämää määräämiskäsitettä käsitteenä, joka hyvin selittää pronominien viittaussuhteita. Heidän mukaansa lukuisat esimerkit osoittavat sen selitysvoiman. Kuno (1987) kuitenkin kritisoi sääntöä. Hänen mukaansa seuraavan virkkeen lausekkeet *John* ja *he* sääntö selittää eriviitteisiksi, mitä ne eivät kuitenkaan ole.

19. Near the girl John was talking with,
he found a snake.
(Kuno 1987)

Reinhartin (1983) mukaan tietynlaiset syntaktiset rakenteet ovat valikoutuneet luonnollisiin kieliin sen vuoksi, että ne ovat kognitiivisesti helposti prosessoitavissa. Tämän Reinhartin esittämän hypoteesin mukaan syntaktiset rakenteet hei jastavat joitakin mielen perinnöllisiä ominaisuuksia.

Syntaktisiin rajoitteisiin perustuvilla menetelmillä on merkitystä automaattisessa anaforan avaamisessa, mutta yksinään ne eivät ole riittäviä (Hirst 1981). Syntaktiset säännöt selittävät pronominin suhteita vain lauseen tai virkkeen sisällä, eikä niistä ole apua virkkeiden välisissä viittaussuhteissa (Grishman 1986). Lisäksi on olemassa melko paljon aineistoa,joka on ristiriidassa esitettyjen sääntöjen kanssa (Kuno 1987).Seuren (1985) ei pidä esitettyjä sääntöjä kovinkaan hyvinä. Hänen mukaansa syntaktisia rajoitteita on olemassa, mutta ainakaan toistaiseksi kukaan ei ole pystynyt niitä täsmällisesti formuloimaan.

Hobbs (1978) kehitti algoritmin, joka otti huomioon suvun, luvun ja kaksi ensimmäistä edellä esitetyistä syntaktisista rajoitteista (Lees ja Klima 1963 sekä Langacker 1969) ja johon sisältyi sanan merkitykset huomioonottavia eli valinnaisia rajoitteita (selectional constraints; ks. seuraava sivu.). Algoritmin testauksessa hänellä oli aineistona arkeologian kirja, Arthur Haleyn romaani sekä Newsweekin artikkeleita.

Valinnaisten rajoitteiden kanssa oikean tuloksen algoritmi antoi 92 %:ssa ja ilman niitä 88 %:ssa tapauksista. Usein kuitenkin oli vain yksi mahdollinen korrelaatti. Kun vaihtoehtoja oli, saatiin oikea tulos valinnaisten rajoitteiden kanssa 82 %:ssa tapauksista. Hobbs ei kuitenkaan lähemmin selitä, mitä hän tarkoittaa "vaihtoehdottomuudella". Itse asiassahan kaikki pronominin edellä olevat oikeaan kategoriaan kuuluvat sana tvoivat olla vaihtoehtoja. Jollei algoritmiin olisi sisältynyt Langackerin esittämää sääntöä, anaforan korrelaatti olisi löytynyt selvästi harvemmin.

Fraurudin (1988) kehittämä algoritmi oli yksinkertaisempi.Siinä käsiteltiin vain yksikkömuotoisia pronomineja eikä siinäotettu huomioon semanttisia rajoitteita (ks. seuraava sivu..). Algoritmissa otettiin huomioon suku, luku ja se, oliko viittauksen kohteena henkilö vai objekti. Siinä otettiin huomioon myös pronominin ja sen korrelaatin välinen etäisyys samoin kuin korrelaatin lauseenjäsenasema. Algoritmiin sisältyi edellä esitelty Leen ja Kliman esittämä sääntö, ja siinä otettiin huomioon Reinhartin kuvaama määräämissääntö. Fraurud sovelsi algoritmia aineistoon, joka koostui kertomuksista,raporteista ja teknisistä artikkeleista. Tulokset on taulukossa 2.

Koko aineistosta algoritmilla löytyi oikea korrelaatti 91 %:ssa tapauksista. Kertomuksissa

		Kertomukset	Raportit	Artikkelit	Yhteensä
Oikein	N	162	166	87	415
	%	99.3	93.3	75.0	90.8
Väärin	N	1	12	29	42
	%	0.7	6.7	25.0	9.52
Yhteensä	N	163	178	116	457
	%	100.0	100.0	100.0	100.0

Taulukko 2. Algoritmin toimivuus: oikean ja väärän korrelaatin jakauma (Fraurud 1988)

algoritmi valitsi väärin vain kerran. Raportteihin sovellettuna se antoi myös suhteellisen hyvän tuloksen, sillä vain n. 7 %:a valinnoista osui harhaan. Sitä vastoin artikkeleissa algoritmin toimivuus oli heikkoa, ja se valitsi väärin joka neljännen korrelaatin. Huomionarvoista on se, että kertomuksissa ja raporteissa oli paljon henkilöpronomineja. Artikkeleissa sitä vastoin ei-henkilöpronominit vallitsivat. (Ks. s. 123–124.) Fraurudin mukaan luvuista ei voi kuitenkaan päätellä sitä, kuinka algoritmi pystyi ratkomaan monitulkintaisuuksia, sillä useissa tapauksissa korrelaattivaihto ehtoja ei "lähietäisyydellä" ollut. Fraurud huomauttaa, että mikä tahansa edellä oleva sopiva sana voi olla korrelaattivaihtoehto, eikä hänellä näin ollen ollut esittää yksiselitteistä tulosta algoritmin kyvystä selvittää monitulkintaisuuksia.

Semanttinen analyysi: semanttiset rajoitteet

Anaforan avaaminen semanttisen analyysin avulla perustuu semanttisten rajoitteiden käyttöön. Rajoitteiden avulla katsotaan, mitkä vaihtoehtoisista korrelaateista olisivat mielekkäitä anaforan paikalla. Tärkeimpiä semanttisia rajoitteita ovat valinnaiset rajoitteet ja predikaatin ala. Sääntö, jonka mukaan verbi vaatii elollisen subjektin, on valinnainen rajoite. Predikaatin ala on mahdollista esittää muuttamalla lauseet loogiseen muotoon. Lause on mielekäs vain, jos predikaatin argumentti kuuluu sen alaan. Esim. predikaatin "on vihreä" ala on kaikki konkreettiset objektit. (Grishman 1986.)

Anafora seuraavassa esimerkissä saataisiin avattua, jos luonnollista kieltä prosessoivassa järjestelmässä olisi esitettynä tieto, jonka mukaan verbin *fly* subjektina voi olla lentokone, muttei laiva.

20. The new flighter took off from Kennedy.
It flew beautifully.
(Grishman 1986)

Rajatuilla aihealueilla valinnaiset rajoitteet ovat terävämpiä ja niiden avulla on mahdollisuus tehdä hienompia eroja, kuin jos niitä käytettäisiin kieleen yleensä. Koko kielen alueella rajoitteiden käyttö on vaikeampaa myös siksi, että kuvitteellinen kielenkäyttö rikkoo rajoitteita. (Grishman 1986.) Lisäksi poikkeuksellisissa tapahtumissa, jollaista kuvaa esimerkki 21, rajoitteet johtavat harhaan.

21. She stared in disbelief at the water coming out of the tap: it was black.
(Hirst 1981)

Tieteellisissä ja teknisissä teksteissä metaforat ja muu kuvitteellinen kielenkäyttö ei yleensä tule kyseeseen. Suppeilla aloilla rajoitteiden käyttö on tehokkaampaa kuin koko kielessä myös sen vuoksi, että suhteessa koko sanastoon ja käsitteistöön, rajoitteita on suppeilla aloilla mahdollista muodostaa enemmän kuin koko kielessä. On kyseenalaista, onko predikaatin ala niin tehokas rajoite, että se soveltuisi kaikkiin tekstityyppeihin. Kielessä kokonaisuudessaan käsitteiden määrä on suuri ja predikaatin alat väljemmin määritelty kuin alakielissä. (Grishman 1986.)

Diskurssin analyysi

Diskurssin analyysissa tarkastellaan diskurssin lauseiden välisiä yhteyksiä ottamalla lingvistisen tekijöiden lisäksi huomioon ekstralingvistiset eli pragmaattiset tekijät. Tilannemerkitys sekä kirjoittajan ja lukijan yhteinen tieto maailmasta vaikuttavat diskurssin rakenteeseen siinä missä syntaktiset ja sanasemanttisetkin tekijät. Diskurssissa sekä

eksplisiittiset tekijät, kuten sanat *nyt, mutta* ja *kuitenkin,* samoin kuin implisiittiset tekijät, kuten tieto maailmasta, liittävät lauseita toisiinsa (Ballard ja Jones 1987). Diskurssin analyysiin perustuva anaforan avaaminen voidaan suorittaa tietämyskuvauksia ja tekstin fokusta hyväksikäyttäen.

Tietämyskuvaukset

Erikois- ja yleistietoa voidaan jäsentää tietokoneen prosessoitavaksi eri tavoilla tiedon sisällön ja "tiedon osasten"keskinäisten suhteiden perusteella. Yksi tietämyksen esitystapa on semanttiset verkot ja toinen kehysesitys. Kehykset kuvaavat jotakin aihekokonaisuutta, esim. keittiötä tai ostoksilla käyntiä, ja niiden sisältö pyritään jäsentämään samalla tavalla kuin ihminen aiheen jäsentää. Kehyksissä on koloja, jotka ovat kehyksen osia tai ominaisuuksia, ja ne voivat saada arvoja, mutta vain tietynlaisia. (Grishman 1986.) Se, mikä ei tekstissä eksplisiittisesti tule esille, voidaan kehysjärjestelmien ja niihin liittyvien sääntöjen avulla päätellä.

Esimerkkinä anaforan avaamiseen vaadittavasta kohdealuetiedosta Carlson ja Honkela (1993) esittävät seuraavan virkkeen.Virkkeessä *se* ei mitä ilmeisemmin viittaa substantiiviin *pöytä.*

22. Lasi tipahti pöydälle ja se meni rikki.

Seuraavat esimerkit havainnollistavat resoluutioon sisältyvää monitulkintaisuutta ja sitä, kuinka tällaisissa tilanteissa oikea korrelaatti voidaan löytää.

23. When Sue went to *Nadia's* home for dinner, *she* served sukiyaki au gratin.
24. When *Sue* went to Nadia's home for dinner, *she* ate sukiyaki au gratin.

(Hirst 1981)

Esimerkissä 23 anafora voitaisiin avata luomalla visiting-kehys. Kehyksiin liittyy odotuksia käsitellystä tekstistä, ja esimerkissä odotuksena on, että *Nadia* saattaisi tarjota *Sue*lle ruokaa. Tämän jälkeen avaaminen olisi helppo päättely. (Hirst 1981.)

Fokuksen merkitys avaamisessa

Sidner (1983) kutsuu fokukseksi sitä diskurssin elementtiä, johon kirjoittaja keskittää huomion. Hirst (1981) katsoo fokusta ja sen suhdetta anaforaan lukijan näkökulmasta: fokus on niiden diskurssin konstituenttien joukko, jotka luettaessa ovat lukijan tietoisuudessa ja jotka ovat mahdollisia anaforan korrelaatteja. Hirst (1981) havainnollistaa fokuksen tarpeellisuutta resoluutiossa seuraavilla esimerkeillä:

25. John left the window and drank the *wine* on the table. *It* was good.
26. John left the window and drank the wine on the table. *It* was brown and round.

Esimerkissä 25 *it* selvästi viittaa sanaan *wine.* Esimerkin 26 lauseista tekee hämärän se, ettei pöytä ole fokuksessa eikä siihen voida anaforisesti viitata. Fokuksessa on viini; Hirstin ja hänen informanttiensa mukaan pronomini *it* viittaa viiniin (Hirstin mukaan viinin yhteydessä on yleistä käyttää "omituisia" adjektiiveja). Todettakoon tässä yhteydessä, että seuraavaksi käsiteltävästä Sidnerin esityksestä voi päätellä, että fokuksessa oleva elementti on hyvin usein lauseen subjekti tai objekti.

Fokus kuuluu olenaisena osana luonnolliseen kieleen. Anaforat viittaavat fokuksessa oleviin konstituentteihin. Jollei anaforan resoluutiojärjestelmä sitä ota huomioon, se antaa virheellisiä vastauksia. (Hirst 1981.)

Fokuksen merkitystä anaforan avaamisessa ovat tutkineet mm.Sidner (1983) ja Webber (1978). Seuraavassa käsitellään Sidnerin tutkimusta. Sidner havainnollistaa fokuksen käsitettä seuraavalla esimerkillä.

27. I want to schedule a *meeting* with Ira.
28. *It* should be at 3 p.m.
29. We can get together in his office.
30. Invite John to come, too.

Lauseen 27 elementti *meeting* on fokus. Lukija ymmärtää, että kirjoittaja puhuu tapaamisesta, sekä lauseessa 28 olevan anaforan ansiosta että yleisen arkitietämyksensä perusteella. Hän tietää, että tapaamisilla on aika, paikka ja osanottajat. Esimerkistä voi huomata, että toisaalta anafora auttaa fokuksen löytämisessä ja toisaalta fokus on anaforan korrelaatti.

Sidnerin ehdottama algoritmi toimii kahdessa vaiheessa. Ensiksi se määrittää eri vaihtoehdoista fokuksen, minkä jälkeen fokuksessa oleva elementti valitaan korrelaatiksi, jos se täyttää syntaktiset, semanttiset ja päättelyyn perustuvat kriteerit.

Fokuksen etsimisessä ja vaihtamisessa käytetään syntaktisia (esimerkit 31 ja 32) ja semanttisia (esimerkit 33 ja 34) indikaattoreita. Syntaktisten indikaattorien käyttö perustuu siihen, että joissakin syntaktisissa lausetyypeissä, kuten esimerkkien 31

ja 32 lohkolauseissa, fokus määrittyy lauserakenteen perusteella.

31. It was *Henrietta* who ate the rutabagas.
32. It was *the rutabagas* that Henrietta ate.

Semanttiset kategoriat indikoivat myös fokusta.

33. Mary took a *nickel* from her toy bank yesterday.
34. She put *it* on the table near Bob.

Semanttinen kategoria teema indikoi fokusta. Lauseessa 33 *a nickel* on lauseen teema ja siten fokus, ja lauseen 34 pronomini viittaa siihen. Teema on objektia vastaava lauseen elementti, kun luokittelu perustuu verbin semanttisiin kategorioihin.

Sidnerin algoritmiin sisältyy enemmän tekijöitä kuin tässä on mahdollista ottaa esille. Kaksi päätekijää ovat diskurssi- ja toimijafokukset. Diskurssifokus on pääfokus, puheena oleva asia. Toimijafokus on agentti, tekijä jossakin tapahtumassa. Agentin käsite jää esityksessä epäselväksi. Kyse on ilmeisesti subjektia vastaavavasta lauseen elementistä, kun elementtejä määritetään semanttisin perustein. Pääsääntöisesti korrelaatit etsitään seuraavasti. Jos pronomini on lauseessa muussa kuin agenttiasemassa, valitaan diskurssifokus korrelaatiksi edellyttäen, ettei mikään syntaktinen, semanttinen tai päättelyyn perustuva rajoite sitä hylkää. Diskurssifokuksen tunnistaminen perustuu edellä esitettyihin syntaktisiin ja semanttisiin indikaattoreihin. Jos pronomini on agenttiasemassa, valitaan korrelaatiksi toimijafokus, ja otetaan huomioon samat ehdot kuin edellä. Jos esim. jokin semanttinen rajoite hylkää diskurssi- ja agenttifokuksen, anaforan korrelaatiksi valitaan joku muu muista vaihtoehdoista.

Tämän pääsäännön avulla löydetään seuraavan esimerkin pronomineille *they* ja *it* korrelaatit. Esimerkissä *Alfred* ja *Zohar* ovat toimijafokus ja *baseball* diskurssifokus.

35. *Alfred* and *Zohar* liked to play *baseball. They* played *it* everyday after school before dinner.

Sidner tarkastelee joitakin esimerkkejä, joissa hänen fokukseen perustuva avaamismallinsa ontuu. Erityisen ongelmalliseksi hän näkee seuraavantyyppiset paralleeliset rakenteet.

36. The green Whitierleaf is most commonly found near *the wild rose.*
37. The wild violet is found near *it* too.

Jotta avaaminen onnistuisi, algoritmin tulisi päättelyn perusteella hylätä väärä fokus, *Whitierleaf.* Tässä tapauksessa tällainen päättely ei kuitenkaan yleisen tiedon avulla onnistuisi. Fokus ei Sidnerin mukaan auta paralleelisten rakenteidenavaamisessa, vaan diskurssista on löydettävä jokin toinen mekanismi niiden selittämiseksi.

6. Anaforat tiedonhaun tutkimuksen ongelmana

Anaforien tunnistaminen on tarpeen mm. kun mekaanisesti luodaan abstrakteja. Tällainen järjestelmä tunnistaa ja erottaa dokumenteista abstrakteihin sisällytettäviä informatiivisia virkkeitä. Jotta saataisiin yhtenäisistä tekstijaksoista rakentuvia abstrakteja, potentiaalisten anaforien joukosta on kyettävä löytämään todelliset anaforat ja näistä vielä sellaiset, jotka viittaavat edellisten virkkeiden sanoihin. Varhaisimmista tietokoneilla tuotetuista "abstrakteista" puuttui sidoksisuus. Dokumenteista erotettiin informatiivisia virkkeitä, joihin anaforat jäivät ikään kuin roikkumaan. Kyse oli pikemminkin ekstrakteista kuin abstrakteista. Lancasterin yliopistossa on kehitelty tietokoneohjelmaa, joka paitsi tunnistaa anaforia myös arvioi, onko korrelaatti eli viittauksen kohde samassa virkkeessä vai muualla kuin anafora. (Paice 1990.)

Syracusen yliopistossa on tehty laaja tutkimus anaforien vaikutuksesta tiedonhakuun. Tutkimuksen tärkeimpiä ongelma-alueita olivat: anaforien luokittelu; sääntöjen kehittäminen funktionaalisten, todellisten anaforien erottamiseksi potentiaalisista anaforista; anaforien taajuus tieteellisten ju kaisujen abstrakteissa; avaamisen vaikutus eri termipainokaavoihin; avaamisen yhteys relevanssiin. (Liddy ym. 1990.) Luokittelua ja taajuutta (s. 123) on käsitelty edellä. Seuraavassa tarkastellaan tutkimuksen antamia vastauksia muihin asetettuihin kysymyksiin. Tutkimusaineisto koostui PsycINFO- ja INSPEC-tietokantojen abstrakteista.

Osa tekstin sanoista on potentiaalisia ja osa funktionaalisia anaforia. Potentiaaliset anaforat ovat sanoja, jotka voivat toimia anaforina, ja funktionaaliset anaforat ovat anaforia tarkasteltavassa tekstinkohdassa. Tutkijat kehittivät leksikaaliseen ja syntaktiseen informaatioon perustuvat säännöt, joiden avulla funktionaaliset anaforat voidaan tunnistaa potentiaalisten anaforien joukosta. Tarkoituksena oli eksplikoida ne prosessit, joiden avulla ihminen tunnistaa, milloin sana on anafora. Hyvät säännöt ovat perustana algoritmeille. Alla on esitetty, kuinka usein säännöt toimivat kussakin anaforaluokassa (ks.tästä luokittelusta s. 123). Toimi-

vuuden kriteerinä oli, että kolme tuomaria sovelsi sääntöä oikein. (Liddy ym. 1987.)

1.	Pronominit	98 %
2.	Demonstratiivipronominit	87 %
3.	Relatiivipronominit	93 %
4.	Nominin substituutit	88 %
5.	Apuverbi	99 %
6.	Indefiniittipronominit	89 %
7.	Adjektiivit	86 %
8.	Adverbiaalit	96 %
9.	Substantiivit	83 %

Liddy ym. (1987) toteavat sääntöjen toimivuuden osoittavan,että algoritmien kehittäminen on mahdollista. Jos lisäksi käytettäisiin semanttista informaatiota, virheiden määrä vähenisi.

Bonzi ja Liddy (1989) esittivät hypoteesin, jonka mukaan anaforat viittaavat keskeisin käsitteisiin tieteellisissä abstrakteissa. Hypoteesi todettiin oikeaksi. Kuitenkin sekä anaforaluokkien että tutkimustietokantojen välillä oli eroja. Esim. demonstratiivipronomit viittaavat hyvin usein keskeisiin ja vain harvoin perifeerisiin käsitteisiin, kun taas relatiivipronominit edustavat usein molempia. INSPECin ja PsycINFOn välillä on myös ero, ja edellisessä anaforat viittaavat useammin keskeisiin käsitteisiin.

Bonzi ja Liddy (1989) tutkivat myös, mikä vaikutus eri termipainokaavoilla on avaamisen ansiosta muuttuviin termien painoarvoihin. He huomasivat, että ne kaavat, jotka eivät ota huomioon dokumenttien pituutta, lisäävät avattavien termien painoarvoja huomattavasti enemmän kuin ne kaavat, joihin dokumenttien pituus sisältyy tasapainoittavana tekijänä. Painoarvojen muuttumiseen vaikuttaa myös se, mistä anaforaluokasta on kyse. Bonzin ja Liddyn (1989) mukaan sen perusteella, että anaforat viittaavat keskeisiin käsitteisiin ja että termien painoarvot kasvavat avaamisen ansiosta, voidaan odottaa, että avaaminen parantaa hakutulosta.

Avaaminen ei kuitenkaan auta erottamaan relevantteja dokumentteja epärelevanteista, eikä sen ansiosta relevanssijärjestys parane. Hakijan kysymyksessä esiintyvillä sanoilla on nimittäin jo ennen avausta suurempi painoarvo kuin muilla sanoilla.Termin keskeisyys ei myöskään korreloi kuin hyvin vähäisessä määrin avaamisen ansiosta saavutettavan termin painoarvon kasvun kanssa. Näiden tulosten mukaan avaamisella ei olisi vaikutusta hakutulokseen. (Bonzi ja Liddy 1989.) Liddy (1990) toteaa, ettei Syracusen yliopistossa tehty tutkimus antanut selvää vastausta anaforan merkityksestä tiedonhaussa. Tärkeä havainto on kuitenkin, että eri anaforaluokilla on erilainen vaikutus (Bonzi ja Liddy 1989). Liddy (1990) tähdentää, että anafora ei ole erillinen ilmiö diskurssissa, ja näin ollen sen ymmärtäminen edellyttää muiden diskurssin ilmiöiden ymmärtämistä ja huomioonottamista. Tätä kautta on mahdollisuus kehittää tehokkaampia hakujärjestelmiä.

Pro gradu -työssäni (Pirkola 1994) halusin ensisijaisesti selvittää sen, mikä vaikutus anaforien ja vaillinaisten ilmaisujen eli ellipsien avaamisella olisi läheisyysoperaatioilla tehtävien kokotekstihakujen tulokseen. Tutkimustietokantana oli Tampereen yliopiston Informaatiotutkimuksen laitoksen sanomalehtitietokanta, jota käytettiin Topic-ohjelman alaisuudessa. Aineisto oli erityyppistä kuin Syracusen yliopiston tutkijoilla, sillä uutistietokannan juttu on keskimäärin pitempi kuin viitetietokantojen tiivistelmä ja se on luonteeltaan ja rakenteeltaan (yleensä paljon kappaleita) erilainen. Sekä ellipsien että anaforien avaaminen osoittautui hyödylliseksi silloin,kun ellipsin ja anaforan korrelaatti (hakuavain) oli tyypiltään erisniminen sanaliitto (esim. Salman Rushdie, Tampella Power). Avaamisen ansiosta läheisyysoperaatiohakujen saanti lisääntyisi näissä tilanteissa selvästi.

7. Johtopäätökset

Edellä on kielitieteen, tietokonelingvistiikan ja tiedonhaun tutkimuksen kirjallisuuden avulla tarkasteltu anaforan käsitettä, luokittelua, taajuutta tekstissä, resoluutiomenetelmiä ja merkitystä tiedonhaussa. Näin lopuksi yhteenvetona ja johtopäätöksenä voidaan kiinnittää huomiota erityisesti kahteen asiaan. Toinen liittyy tiedonhakujärjestelmien kehittämiseen ja toinen IR-tutkimukseen yleensä.

Voidaan pohtia, mitä mahdollisuuksia on kehittää sellainen luonnollista kieltä käsittelevä järjestelmä, joka onnistuneesti pystyisi tekemään anaforan resoluution. Tietyissä teksti-, anafora- ja korrelaattityypeissä ilmeisestikin jo suhteellisen yksinkertaisella menetelmällä on mahdollista saavuttaa hyvä tulos avaamisessa. Jollei mitään rajoituksia aseteta, niin anaforaa ei kuitenkaan voi tulkita eikä sitä voi avata, jollei oteta huomioon kielen kaikkia tasoja, siis morfologiaa, syntaksia, semantiikkaa ja pragmatiikkaa. Mitä ylemmäksi siirry-

tään, sitä vaikeampaa avaaminen on. Syntaktisen ja osittain semanttisen tason avaaminen voi nojautua kielen yleispäteviin käyttösääntöihin. Pragmaattisella tasolla sitä vastoin järjestelmän tietämyskannassa tulee olla esitettynä arkitietämystä ja järjestelmän tulee pystyä päättelemään. Näin ollen sellaisen resoluutioalgoritmin kehittäminen, joka soveltuisi kaikkeen kirjoitettuun tekstiin, on äärimmäisen vaikeaa. Fraurudin (1988) mukaan se on mahdotonta, jos tavoitteena on 100 %:n tulos. Sellaisen järjestelmän kehittäminen, joka "ymmärtäisi" koko New York Times -lehden numeron, edellyttäisi hyvin pitkälle kehittyneen lingvistisen tekniikan hyväksikäyttöä, ja järjestelmän tulisi sisältää tietämyskanta, joka pystyisi käsittelemään kaikki esiintulevat aiheet (Rich 1987). Rajatuilla alueilla kuitenkin se informaatio, joka tarvitaan hyvin toimivien resoluutioalgoritmien kehittämiseksi, on hallittavissa, ja tietyissä tilanteissa tehokkaiden resoluutiojärjestelmien kehittäminen vaikuttaa mahdolliselta.

Toinen yleisemmän tason johtopäätös, joka käsitellystä kirjallisuudesta voidaan tehdä on se, että informaatiotukimuksen tulee olla läheisessä yhteydessä lingvistiikkaan. Siinä prosessissa, jossa informaatiota haetaan tietokannasta, yhdellä puolella on tiedonhakija tiedontarpeineen. Välissä on tietokanta ja toisella puolella informaation tuottaja. Tuottaja esittää sanottavansa kirjoitetulla luonnollisella kielellä, informaatio esitetään tietokannassa sitä käyttäen ja se on myös tiedonhakukieli. Alussa käsiteltiin sitä, mikä alan tutkimuksen perimmäiseksi tavoitteeksi on asetettu. Kerrattakoon se vielä: tiedon tallennuksen ja haun tutkimuksen perimmäinen tavoite on kehittää käsitteitä, menetelmiä ja järjestelmiä, joiden avulla kaikki tieto, olipa se missä tahansa muodossa ja missä tahansa paikassa, saadaan vaivattomasti kenen tahansa sitä tarvitsevan ulottuville ja siten esitettynä, että tarvitsijan on mahdollisimman helppo tämä tieto omaksua. Kun tavoite on tämä ja kun tiedonhakuprosessin kaikissa vaiheissa käytetään luonnollista kieltä, niin on selvää, että IR-tutkimuksen on liityttävä lingvistiseen tutkimukseen. Tällaista lingvististä IR-tutkimusta harjoitetaankin. Anaforan tutkimus on tästä hyvä esimerkki. Tutkimus on ollut hedelmällistä ja askel kohti tiedon tallennuksen ja haun tutkimuksen perimmäistä tavoitetta.

Hyväksytty julkaistavaksi 22.11.1994.

Lähdeluettelo

Akmajian, A. & Demers, R. & Farmer, A. & Harnish, R. 1990.Linguistics: An Introduction to Language and Communication.Cambridge, Mass.: The MIT Press.

Ballard, B. & Jones, M. 1987. Computational linguistics.Teoksessa: S. Shapiro (toim.) Encyclopedia of Artificial Intelligence. John Wiley & Sons, 133–151.

Bonzi, S. & Liddy, E. 1989. The Use of Anaphoric Resolution for Document Description in Information Retrieval. Information Processing & Managenent 25(4), 429–441.

Carlson, L. & Honkela, T. 1993. Luonnollisen kielen käsittely.Teoksessa: E. Hyvönen, I. Karanta, M. Syrjänen (toim.) Tekoälyn ensyklopedia. Hämeenlinna: Karisto, 233–243.

Carter, D. 1987. Interpreting Anaphors in Natural Language Texts. New York: Wiley & Sons.

Deemter, K.van. 1992. Towards a Generalization of Anaphora.Journal of Semantics 9(1), 27–50.

Fraurud, K. 1988. Pronoun Resolution in Unrestricted Text.Nordic Journal of Linguistics 11(1–2), 47–68.

Grishman, R. 1986. Computational linguistics: An introduction.New York: Cambridge University Press.

Hakulinen, A. & Karlsson, F. 1988. Nykysuomen lauseoppia.Jyväskylä: Gummerus.

Halliday, M. & Hasan, R. 1976. Cohesion in English. London:Longman.

Hirst, G. 1981. Anaphora in Natural Language Understanding:A Survey. Lecture Notes in Computer Science 119. Berlin: Springer-Verlag.

Hobbs, J. 1978. Resolving Pronoun References. Lingua 44,311–338.

Jackendoff, R. 1972. Semantic Interpretation in Generative Grammar. Cambridge, Mass.: The MIT Press.

Järvelin, K. 1992. Tiedonhaun moniste. Tampereen yliopisto.Kirjastotieteen ja informatiikan laitos.

Karlsson, F. 1982. Johdatusta yleiseen kielitieteeseen. Helsinki: Gaudeamus.

Kuno, S. 1987. Functional Syntax: Anaphora, Discourse and Empathy. Chicago: The University of Chicago Press.

Langacker, R. 1969. On Pronominalization and the Chain of Command. Teoksessa: D. Reibel, S. Schane (toim.) Modern Studies in English: Readings in

Transformational Grammar. New Jersey: Prentice-Hall, 160–186.
Lees, R. & Klima, E. 1963. Rules for English pronominalization.Language 39(1), 17–28.
Liddy, E. 1990. Anaphora in Natural Language Processing and Information Retrieval. Information Processing & Management26(1), 39–52.
Liddy, E. & Bonzi, S. & Katzer, J. & Oddy, E. 1987. A Study of Discourse Anaphora in Scientific Abstracts. Journal of the American Society for Information Science 38(4), 255–261.
Paice, C. 1990. Constructing Literature Abstracts by Computer:Techniques and Prospects. Information Processing & Management26(1), 171–186.
Pirkola, A. 1994. Ellipsien ja anaforien taajuus tekstissä ja niiden avaamisen vaikutus läheisyysoperaatioilla tehtävissä tiedonhauissa. Informaatiotutkimuksen pro gradu -tutkielma. Tampereen yliopisto. Informaatiotutkimuksen laitos.
Reinhart, T. 1983. Anaphora and Semantic Interpretation. London: Croom Helm.
Rich, E. 1987. Artificial Intelligence. Teoksessa: S. Shapiro (toim.) Encyclopedia of Artificial Intelligence. John Wiley &Sons, 9–16.
Seuren, P. 1985. Discourse Semantics. Oxford: Basil Blackwell.
Sidner, C. 1983. Focusing in the Comprehension of Definite Anaphora. Teoksessa: M. Brady, R. Berwick (toim.) Computational Models of Discourse. Cambridge, Mass.: The MIT Press, 267–330.
Webber, B. 1978. A Formal Approach to Discourse Anaphora.New York: Garland Publ.

Termit ja Merkinnät

Dominoida – Puurakenteessa ylempänä oleva solmu dominoi alempia solmuja.

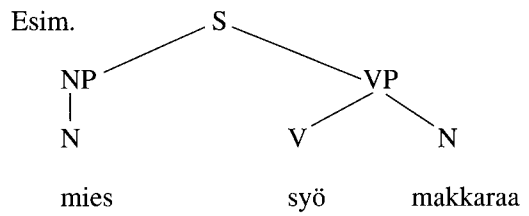
Intensio – Ne asiat eli tunnusmerkit, jotka olennaisesti kuuluvat käsitteeseen.

Konstituentti (muodostin) – Jokainen morfeemi, sana tai konstruktio, joka kuuluu osana laajempaan konstruktioon.

NP (noun phrase) – Nominaalilauseke, substantiivilauseke.

Parafraasi – Jonkin ilmauksen merkityksen esittäminen toisin sanoin.

Puurakenne – Puurakennetta käytetään syntaktisten suhteiden kuvaamiseen. Syntaktisesta puusta ilmenee lauseen konstituenttien dominointisuhteen lisäksi konstituenttien keskinäinen järjestys sekä niiden kieliopillinen kategoria (kullakin solmulla on oma nimekkeensä).



S (sentence) – S-solmu eli lausesolmu.

Samaviitteinen – kaksi ilmausta ovat samaviitteiset, kun niiden tarkoite on identtinen. Esim. poika ja hän seuraavassa virkkeessä: Poika kehuskeli, että hän osasi lumota käärmeitä.

Solmu – Puurakenteen nivelkohta (esim. S, NP, VP jne.).

Tarkoite – Ulkomaailmassa oleva olio tai oloseikka, johon sana viittaa.

VP (verb phrase) – verbilauseke.

Lähteet: Fakta-tietosanakirja (1971) ja A. Hakulinen & J. Ojanen: Kielitieteen ja fonetiikan termistöä (1976).