RIITTA NURMINEN

Suomen kielen automaattinen analyysi ja sen hyödyntäminen tiedonhaussa

Nurminen, Riitta, Suomen kielen automaattinen analyysi ja sen hyödyntäminen tiedonhaussa [The automatic analysis of the Finnish language and its use in information retrieval]. Kirjastotiede ja informatiikka 5(2): 56—62, 1986.

The article deals with the effects of the specific features of the Finnish language in information retrieval. The author describes the projects aiming at automatic interpretation of the Finnish language and four programs which automatically analyze wordforms and generate their stem variants or basic forms. The tests conducted with these programs in the context of the MINTTU information retrieval system are described. Finally, six alternative system architechtures are outlined.

Address: Technical Research Centre of Finland, Information Centre, Vuorimiehentie 5, SF-02150 Espoo, Finland.

Useimmat tiedonhakuun liittyvät ongelmat ovat yhteisiä kaikille tiedonhakujärjestelmille riippumatta siitä, ovatko kuvailutermit luonnollisen kielen sanoja, poimittu tesuruksesta vai numeerisia koodeja. Eri ihmiset näet liittävät samoihin ilmauksiin erilaisia merkityssisältöjä muun muassa kokemustensa ja koulutuksensa perusteella. Tällaisia hakutermien merkityksen ja kattavuuden ongelmia, sisällöllisiä ongelmia, esiintyy tiedonhaussa aina. Automaattisissa tiedonhakujärjestelmissä ne vain ovat erityisen hankalia sen vuoksi, koska järjestelmät eivät pysty luovaan tulkintaan vaan perustuvat käyttäjän antaman hakutermin ja käänteistiedoston kuvailutermin ehdottomaan täsmäävyyteen.

Vaikka merkitykseen liittyvät ongelmat ovatkin olennaisimpia tiedonhaussa, vaikuttaa myös termien muoto tiedonhaun onnistumiseen. Vapaatekstihaussa järjestelmän käyttäjän on osattava kielioppinsa, jotta typistetty muoto palauttaa hakusanan kaikki eri taivutusmuodot. Jos hakutermi typistetään väärästä paikasta, voi jouko asiaankuuluvia dokumentteja jäädä löytämättä. Vaikka kirjastonhoitajat ja informaatikot kokemuksen kautta oppivatkin hallitsemaan sanojen taivutuksen, voi kokenutkin tiedonhakija epähuomiossa tehdä väärän ratkaisun. Ongelmasta tulee vielä visaisempi, jos tiedontarvitsija itse tekee haun, sillä maallikon on vielä vaikeampi pitää mielessään kaikkia tiedonhakuun liittyviä yksityiskohtia.

Tieteellisen informoinnin neuvoston ja Valtion teknillisen tutkimuskeskuksen informaatiopalvelulaitoksen rahoittamassa tutkimuksessa selvitettiin, voidaanko nykyisissä suomalaisissa tiedonhakujärjestelmissä paremmin ottaa huomioon suomen kielelle ominaisia piirteitä. Tämä voidaan toteuttaa muun muassa liittämällä järjestelmiin suomea automaattisesti analysoivia ohjelmia. Tutkimuksessa vertailtiin ohjelmia, jotka tuottavat sanoista taivutusvartaloita eli sanan taivutusmuodoille yhteisen kantaosan, sekä ohjelmia, jotka palauttavat perusmuotoon niille syötetyn saneen eli sanan taivutusmuodossa olevan esiintymän. (Nurminen 1986.)

Suomen kielen erityispiirteet

Tekstitiedonhaun ja -hallinnan, kuten yleensäkin automaattisen tietojenkäsittelyn, valtakieli on englanti. Englannin kieliopillinen rakenne kuitenkin poikkeaa sen verran suomen kielestä, että ulkomailla kehiteltyjä ohjelmistoja ei voida ongelmitta suoraan soveltaa suomenkieliseen tekstiin. Esimerkiksi taivutuspäätteitä ja johtimia englannissa käytetään huomattavasti vähemmän kuin suomessa, jossa yksistään sijamuotoja on 14.

Jos lasketaan kaikki ne erilaiset taivutusmuodot, jotka suomen taivutuspäätteiden, omistuspäätteiden ja muiden vastaavien avulla voidaan muodostaa, on suomen kielessä teoriassa mahdollista muodostaa substantiiveista noin 2000, adjektiiveista 6000 ja verbeistä 12 000 erilaista taivutusmuotoa (Koskenniemi 1985a, s. 20).

Periaatteessa suomen kielen saneitten pääteainekset voitaisiin typistää kuten englannissakin siten, että sanasta karsitaan pois kaikki mahdolliset pääteainekset, jolloin saadaan kaikille taivutusmuodoille yhteinen sanavartalo. Tämä ei kuitenkaan riitä, koska suomen kielessä sanojen vartalotkin muuntelevat enemmän kuin englannissa. Vartalot voivat poiketa toisistaan huomattavastikin, kuten *yksi—yhtenä* tai *yötä—öitä*.

Vapaatekstihaussa yhdyssanat ovat erityinen ongelma. Käytännössä yhdyssanojen jälkiosat ovat tiedonhaun ulottumattomissa, ellei tiedonhakujärjestelmä salli hakutermin vasemman puolen katkaisua tai ellei tiedonhakija itse keksi kaikkia mahdollisia sanoja, jotka on liitetty määriteosaksi yhdyssanan alkuun.

Suomen kielen tulkinnan projektit

Suomessa on nykyään käynnissä kaksi eri projektia, joissa pyritään rakentamaan suomen kieltä tulkitsevia ohjelmia.

Helsingin yliopiston yleisen kielitieteen laitoksella on käynnissä suomen kielen automaattisen analysoinnin projekti, jonka pyrkimyksenä on laatia kielitieteellisesti pätevä luonnollisen kielen tietokonemalli.

Alkuvaiheessa tutkimus keskittyi lähinnä sanamuotojen analyysiin eli morfologiaan. Projektin tuloksena on syntynyt monia erilaisia sanoja analysoivia ja muokkaavia ohjelmistoja. Tällä hetkellä tutkimuksen alla on suomen kielen lauserakenteiden automaattinen analyysi. (Karlsson 1985a ja 1985b)

SITRAn Kielikone-projektissa pyritään rakentamaan suomen kieltä ymmärtävä tietokantaliitäntä. Ideana on, että tiedontarvitsijan ei tarvitse erikseen opetella formaalia tietokannan kyselykieltä, vaan tämän ohjelmiston avulla voidaan etsiä tietoa käyttäen suomen kieltä. Samalla on pyritty rakentamaan myös muihin sovellusalueisiin käyttökelpoista ohjelmistoa. (Jäppinen et al. 1985a ja 1985b.)

Projektin rakentama suomen kielen tulkintaohjelmisto jakaantuu neljään moduliin. Kolmessa ensimmäisessä modulissa tehdään tekstin morfologinen, syntaktinen ja semanttinen analyysi; viimeisessä modulissa tämän kieliopillisen analyysin perusteella muodostetaan tietokantakysely.

Hakuvartaloita tuottavat ohjelmat

FINSTEMS on Yleisen kielitieteen laitoksella kehitetty suomen kielen substantiivien taivutusvartaloita tuottava ohjelma. Ohjelman syötteenä on perusmuodossa (yksikön nominatiivissa) oleva substantiivi, esimerkiksi *lapsi*. Tulosteena saadaan syötetyn sanan taivutusvartalot, esimerkiksi *lapsi-, lapse-, last-*.

FINSTEMS ei tarvitse varsinaista sanakirjaa, vaan päättelee sanasta tuotettavat taivutusvartalot sanan kirjoitusasun perusteella. Koska suomen kielen vanhimmat sanat usein poikkeavat normaalisäännöistä, on ne kuitenkin tallennettu omaan tarkistuslistaan, joka käydään läpi ennen muuta analyysiä.

SITRAn kehittelemä TAIVUTIN tuottaa suomenkielisten sanojen taivutusvartaloita. Ohjelman säännöstö kattaa substantiivit, adjektiivit, numeraalit ja verbit. Ohjelman syötteenä on perusmuodossa oleva sana ja tulosteena syötetyn sanan taivutusvartalot.

Eräissä tapauksissa sanan ulkoasu ei ole tarpeeksi yksiselitteinen täsmällisen taivutuksen määrittelemiseen. Jotta kaikki tarvittavat taivutusvartalot tulisivat varmasti lueteltua, tuottavat ohjelmien taivutussäännöt joissain tapauksissa varmuuden vuoksi myös ylimääräisiä vartaloita. Esimerkiksi *rakkaus*-sanalla on vartalo *rakkaude-*, mutta samaan taivutusluokkaan kuuluvan *pakkaus*-sanan muotoa *pakkaude-* ei todellisuudessa esiinny. (Koskenniemi 1985b.)

Saneita perusmuotoon palauttavat ohjelmat

Kaksitasomalli on Yleisen kielitieteen laitoksella laadittu ohjelma, joka pystyy tunnistamaan ja tuottamaan sanojen taivutusmuotoja. Se pystyy myös palauttamaan johdokset kantasanaansa ja purkamaan yhdyssanat osiin.

Syötteenä on suomen kielen sananmuoto eli sane, esimerkiksi *lääkkeitä*. Tuloksena saadaan sanan perusmuoto ja morfologisen analyysin tulos; yllä mainitussa esimerkissä LÄÄKE N PTV PL, eli substantiivi, partitiivi, monikko. Nomineilla perusmuoto on yksikön nominatiivi ja verbeillä

aktiivin indikatiivin preesensin vahva vokaalivartalo (jonka saa poistamalla IV infinitiivin muodosta päätteen *-minen* : *tekeminen* → *teke-*). Yhden saneen analysointiin kuluu aikaa suunnilleen 0.1 sekuntia Burroughs B7800-järjestelmän CPU-aikaa (Koskenniemi 1983, s. 137).

MORFO on yksi SITRAn Kielikone-projektin ohjelmamoduleista, jota käytetään sanamuotojen morfologiseen analyysiin. MORFO pystyy palauttamaan suomenkielisten sanojen taivutetuista muodoista perusmuodot, ja yhdyssanoista niiden osien perusmuodot.

MORFOn syötteenä on sananmuoto, esimerkiksi *rajoituksen*. Analyysin tuloksena saadaan sanan perusmuoto, joka nomineilla on yksikön nominatiivi ja verbeillä I infinitiivi (esimerkiksi *tulla, juosta*). Tulostettavat tiedot riippuvat valitusta MORFON toimintatavasta: sen voi esimerkiksi ohjata tulostamaan vain luettelon niistä sanamuodoista, joita se ei pystynyt analysoimaan.

MORFOn sanakirja kattaa noin 23 000 sanaa. Sanomalehtitekstiä analysoitaessa kuluu yhden saneen analyysiin 60—70 ms VAX-11/750-järjestelmän CPU-aikaa. Ohjelman sanakirja on avoin eli käyttäjä pystyy suhteellisen yksinkertaisesti itse lisäämään tarvitsemiaan termejä sanastoon erityisen valikon avulla.

Koska analyysi tehdään yhden irrallisen sanamuodon perusteella, ohjelmat voivat löytää useita mahdollisia perusmuotoja. Esimerkiksi sane *ilmaisin* voi olla joko substantiivin nominatiivimuoto (genetiivi olisi *ilmaisimen*) tai verbin yksikön enimmäisen persoonan imperfekti (I infinitiivissä muoto *ilmaista*). Se, miten hyvin tällaisia kirjoitusasultaan samankaltaisia eli homografisia muotoja pystytään havaitsemaan, riippuu sanaston koosta. Mitä enemmän sanastossa on sanoja, sitä paremmin pystytään löytämään kaikki mahdollisuudet. Yksiselitteiseen tulkintaan ei aina pelkän morfologisen analyysin pohjalta päästä, vaan syvempi tulkinta vaatii tekstiyhteyden tuntemista.

Vartaloita tuottavien ohjelmien testaukset

Testausympäristöksi valittiin julkisia tietokantoja sisältävä MINTTU-tiedonhakujärjestelmä, koska siinä on mahdollisuus vapaatekstihakuun. Siitä otettiin kolme tietokantaa: KATI, VTT:n tutkimusrekisteri VTT TUVU ja FINLEXistä FKHO.

Testausaineistona käytettiin aitoja hakutermejä, joita oli saatu muun muassa Espoon kaupunginkirjastossa ja VTT:n informaatiopalvelussa tehtyjen hakujen listauksista.

Aluksi FINSTEMS- ja TAIVUTIN-ohjelmia yksinään testattiin joukolla mutkikkaasti taipuvia tai muuten hankalia suomen kielen sanoja, joiden kautta saatiin yleiskuva ohjelmien taivutussääntöjen kattavuudesta (vrt. Karlsson 1983) sekä joukolla tutkimusta varten kerätyistä hakutermeistä.

Kerätyistä tiedonhakulistauksista poimittiin joukko kyselyitä, jotka tehtiin MINTUlla uudelleen. Muutamista hauista mitattiin myös CPU-aika, joka kului järjestelmän saatua etsittävät hakutermit ja ilmoitettua, montako niihin liittyvää dokumenttia käänteistiedostosta löytyi (aikaan ei siis sisälly itse viitetietueiden poimiminen ja tulostaminen dokumenttitiedostosta).

Seuraavassa vaiheessa muodostettiin tiedonhauissa käytettyjen hakutermien taivutusvartalot automaattisesti ja haut tehtiin uudelleen näillä hakutermeillä.

Vartaloiden automaattisessa tuottamisessa on kaksi hankaluutta: ylimääräisten vartaloiden tuottaminen ja liian vähien vartaloiden tuottaminen. Ylimääräisiä vartaloita ei juuri voi poistaa, koska kaikkien poikkeustenkin mukaantulo on varmistettava. Sen sijaan jonkin vartalon puuttuminen on virhe, koska tällöin tähän muotoon liittyvät dokumentit jäävät löytymättä.

Kun hakuvartalot tuotetaan automaattisesti, tiedonhakuun kuluva aika lisääntyy jonkin verran. Ensinnäkin TAIVUTTIMen tai FINSTEMSin lataaminen tietokoneen muistiin kestää tietyn ajan; lisäksi tulee vartaloiden tuottamiseen kuluva aika. FINSTEMS kulutti VAX-11/750 -järjestelmän CPU-aikaa sanaa kohti noin 1,35 sekuntia ja TAIVUTIN noin 0,07 sekuntia.

Edellämainitut toiminnot vievät kokonaisuuden kannalta aikaa suhteellisen vähän. Eniten ajankäyttöön vaikuttaa se, että hakukomennossa on yleensä useampia (1—5 kappaletta) typistettyjä hakutermejä kuin silloin, kun käyttäjä itse typistää hakutermin. Monien typistettyjen hakutermien ja monien komentojen käyttö vie enemmän aikaa kuin yhden lyhyen termin käyttö yhdessä komennossa.

Kun MINTTU-järjestelmä etsi FINSTEMSin tuottamiin useisiin hakutermeihin täsmääviä termejä käänteistiedostosta, CPU-ajan kulutus oli noin 1,7-kertainen verrattuna alkuperäisten hakutermien etsimiseen.

Tutkimuksessa arvioitiin testikyselyiden ja alkuperäisten kyselyiden saannin ja tarkkuuden välisiä eroja. Esimerkkinä saannin paranemisesta on haku, jossa on etsitty noitia käsittelevää kirjallisuutta. Alkuperäinen kysely oli muodoltaan:

HAE NOITA-!NOITIE-!NOIDAT-!NOIDIL-

Vaikka hakukyselyn laatija on ottanut huomioon *noita*-sanan hankalan taivutuksen, on ter-

meihin livahtanut mukaan myös osia taivutuspäätteistä. Hakuun tuli vastaukseksi kaikkiaan 43 viitettä.

FINSTEMSin tuottamia vartaloita käyttämällä kyselystä tuli:

HAE NOITA-!NOIDA-!NOITI-!NOIDI-

Tällöin vastaukseksi saatiin 58 dokumenttiviitettä eli 15 viitettä edellisten lisäksi, joten edellinen kysely ei ollut aukoton. Tosin muodollisesti sopiva termi voi aina palauttaa sisällöllisesti vääriä viitteitä, esimerkiksi Anni Polvan teoksen »Voi noita miehiä». Kokonaisuuden kannalta katsoen automaattisesti tuotettujen termien käyttö kuitenkin paransi saantia.

Tarkkuus lisääntyy kyselyissä, joissa pitempi hakuvartalo rajaa asiaankuulumattomat viitteet pois. Alunperin kieltolaki-aiheesta tehty haku oli muodoltaan:

HAE KIELTOLA-

Tähän hakuun tuli KATI-tietokannasta vastaukseksi 17 viitettä, joista 15 oli asiaankuuluvia ja 2 käsitteli englannin kielen kieltolauseita. Kun haussa käytettiin FINSTEMSin tuottamia pidempiä vartaloita, kysymys oli:

HAE KIELTOLAKI-!KIELTOLAI-!
KIELTOLAKE-!KIELTOLAE-

Vastaukseksi saatiin 15 viitettä eli samat relevantit viitteet, mutta ei edellämainittuja kahta epärelevanttia viitettä.

Jos järjestelmä tarjoaa mahdollisuuden käyttää automaattisesti tuotettuja vartaloita, käyttäjän on voitava ratkaista, milloin käyttää mahdollisuutta ja milloin typistää termin itse. Pitkät sanat, kuten *liikevaihtoverolaki,* on helppo haun sekaantumatta typistää, esimerkiksi muotoon *liikevaihtoverol-*. Tässä automaattinen taivutus tuottaisi vain neljä eri vartaloa, joita käyttäen saataisiin samat viitteet. Lyhyitä sanoja on kuitenkin vaikea typistää vielä lyhyemmiksi hakusanoiksi, koska tällöin todennäköisesti saadaan runsaasti asiaankuulumattomia viitteitä, eli haun tarkkuus vähenee.

Yleensä automaattisesti tuotettuja termejä käyttäen saatiin samat viitteet kuin alkuperäisissä kyselyissä. Poikkeukset olivat kolmea tyyppiä: Ensinnäkin alkuperäisestä hausta puuttui jokin taivutusmuoto, jolloin automaattisesti tuotetut vartalot löysivät unohdetun muodon; saanti siis parani. Toisaalta alkuperäisessä kyselyssä voitiin typistää hakutermi hyvin lyhyeksi, jolloin myös termille läheiset johdokset tulivat mukaan. Saman tuloksen saaminen automaattisesti tuotetuilla pidemmillä vartaloilla vaatii useampien termien käyttöä. Kolmanneksi, alkuperäisiä termejä pidemmät automaattisesti tuotetut vartalot rajasivat pois epäolennaisia viitteitä eli paransivat haun tarkkuutta.

Perusmuotoihin palauttavien ohjelmien testaus

MORFO- ja kaksitasomalliohjelmille annettiin analysoitavaksi VTT TUVU -tietokannasta poimittuja tietojenkäsittelyalan projekteista kertovia tekstejä. Kun MORFOlla oli käytössä 17 000 sanan sanakirja, se jätti aineiston 2439 saneesta analysoimatta 7,4 %. Nämä jaettiin vielä tarkemmin kolmeen osaan. Oikeinkirjoitusvirheitä, jotka analysoimattomuus paljasti, oli 0,6 % sanoista. Varsinaisten kirjoitusvirheiden osuus analysoiduista saneista oli siis varsin pieni. Sanan puuttuminen ohjelman sanakirjasta aiheutti epäonnistumisen 2,2 %:issa saneista. Loput 4,7 % olivat vierasperäisiä sanoja tai näiden lyhenteitä. Esimerkkinä lyhenteestä, jonka taustalla on jokin vieraskielinen ilmaus on CAD = *computer aided design.*

Vierasperäisten ilmausten käyttö luonnollisesti riippuu aihealasta; tietojenkäsittelyalalle se on erityisen tyypillistä. Vaikka tekstissä puhutaankin suomeksi tietokoneavusteisesta suunnittelusta, lyhenteenä käytetään kuitenkin kirjainjonoa CAD. Varsinainen ongelma onkin, pitäisikö tällaisia vierasperäisiä sanoja lisätä suomen kieltä analysoivan ohjelman sanastoon vai ei — eivähän ne välttämättä taivu suomen kielen sääntöjen mukaisesti.

Tutkitun aineiston perusteella MORFOn kieliopillisessa analyysissä ei juurikaan ole virheitä. Useimmille saneille oli saatu yksi ja oikea perusmuoto. Joskus saneille oli löydetty ylimääräisiäkin tulkintoja, esimerkiksi muodolle *tiedonhallintaan* esitettiin vaihtoehdoksi yhdistelmä joko sanoista *tieto* ja *hallinta* tai sanoista *tieto, halli* ja *taka.* Automaattinen analyysi siis havaitsee monia tulkintoja silloinkin, kun ilmaus ihmisen mielestä on yksitulkintainen. Koska ohjelmat tekevät päätelmänsä vain muodollisin morfologisin perustein, ylitulkintaa ei voida täysin välttää.

Ylitulkinta ei ole varsinainen virhe, joskin se on syytä karsia mahdollisimman vähiin. Virheeksi on laskettava ne tapaukset, kun ohjelmisto antaa sanamuodolle vain väärän tulkinnan. Tällaisia tulkintavirheitä MORFO teki 2439 saneen aineistossa kaksi; väärintulkittujen homografien osuus siis oli 0,08 % sanamuodoista. Näissä tapauksissa ohjelma olettaa tehneensä oikean analyysin, joten käyttäjällä ei ole mitään mahdollisuutta jäljittää virhettä kuten silloin, kun ohjelma tulostaa analysoimattomat sanamuodot näkyville.

Toinen väärintulkituista saneista oli *keskusohjaimen,* jolle MORFO ehdotti tulkinnaksi yhdistelmää sanoista *keskus, ohjat* ja *imeä,* koska sen sanastosta puuttui sana *ohjain.* Toinen virhe tuli muodosta *on-line-tietokonejärjestelmäsuunnitelma,* josta MORFO tulkitsi ensimmäisen sanan *olla*-verbin taivutusmuodoksi. Lisäksi ohjelma harhautui yhdessä kirjoitusvirheessä, jossa eri riveille jaotellun sanan osien väliltä puuttui tavuviiva eli: *tar kastellaan.* Ohjelma tunnisti loppuosan verbiksi *kastella* ja ilmoitti, ettei pysty analysoimaan muotoa *tar,* joten käyttäjä sai tiedon kirjoitusvirheestä. Kirjoitusvirheet tosin paljastuvat vain sillä ehdolla, että virheellinen sane ei ole jonkin muun todellisen sanan esiintymä ja siten tule analysoiduksi.

Kaksitasomallin käytössä oli 10 000 sanan sanakirja ja se jätti 2439 saneesta analysoimatta 18,2 %. Jaoteltuna näistä 0,6 % oli kirjoitusvirheitä ja 4,5 % tapauksista analysoimaton sana oli vierasperäinen sana tai sellaisen lyhenne. 13,2 % tapauksista analyysi estyi, koska sana puuttui ohjelman sanastosta.

Yleensä MORFO pystyi analysoimaan useamman saneen kuin kaksitasomalli, mikä oli lähinnä suuremman sanaston ansiota. Toisaalta Kaksitasomalli analysoi johdokset hieman perusteellisemmin kuin MORFO ja pystyi näin tunnistamaan esimerkiksi muodot *vaikeaselkoisuus, teknistaloudellista* ja *halkeilun,* joista MORFO ei selvinnyt. Vaikka esimerkiksi *halkeilu* pystytäänkin palauttamaan verbivartaloon *halkea-,* ei pitkälle johtava palauttaminen silti ole kaikkien sanojen kohdalla yksinomaan eduksi. Esimerkiksi *muuttujan* palauttaminen verbiin *muutta-* tai *taloudellinen*-muodon palautus *talous*-sanan lisäksi *talo*-sanaan ovat tavallisen tiedontarvitsijan kannalta ylimääräistä kielihistoriaa.

Kaksitasomalli teki saman väärintulkinnan *online*-saneen kohdalla ja jätti *keskusohjaimen* kokonaan analysoimatta. Lisäksi Kaksitasomalli tulkitsi muodon *valaistus* väärin yhdyssanaksi *vala* + *istu-*. Väärintulkittujen homonyymien osuus oli näin 0,12 % saneista.

Perusmuotojen vaikutus hakutermien muotoon

Perusmuodossa olevilla hakutermeillä on se etu, että käyttäjän ei tarvitse vaivata päätään sanojen taivutuksella, vaan hän voi antaa järjestelmälle suoraan sanan perusmuodon. Tiedonhakijan ja järjestelmän käsitys perusmuodoista ei kuitenkaan välttämättä ole sama. Järjestelmä esimerkiksi voi palauttaa takaisin kantasanaansa sellaisia johdoksia, joita tiedonhakija ei välttämättä itse miellä johdoksiksi.

Yleensäkin saneita perusmuotoon palauttavan järjestelmän pitäisi olla johdonmukainen analyysiltään, jotta samantyyppisiä sanoja ei analysoitaisi milloin yhdellä, milloin toisella tavalla. Tavallisen tiedonhakijan on muutenkin päätteellä istuessaan vaikea hahmottaa analyysin taustalla olevaa kielitieteellistä logiikkaa.

Monien sanojen merkitys on etääntynyt varsin kauas siitä sanasta, josta nämä sanat on johdettu, esimerkkinä sanat *maa* ja *maallinen.* Tällaista johdoksen merkityksen itsenäistymistä, jolloin johdos selvästi tarkoittaa eri käsitettä kuin kantasanansa, sanotaan leksikaalistumiseksi. Vaikka käyttäjä pystyykin kielihistoriallisesti mieltämään *teollisuus*-sanan *tehdä*-sanan johdannaiseksi, hän teollisuusrakennuksiin liittyvää tietoa hakiessaan ei varmasti nykyisin käytä *tehdä-* ja *rakentaa*-sanoja hakutermeinä.

Voidaan huomauttaa, että pelkän luonnollisen kielen käyttö sellaisenaan asettaa tiedonhakujärjestelmän käyttäjälle liian suuret vaatimukset (Fugmann 1982, Blair ja Maron 1985). Jos termeistä vain poistetaan sijapäätteet ja liitepartikkelit, lähekkäisten johdosten yms. termien keksiminen vaatii käyttäjältä paljon vaivaa.

Periaatteessa saanti saattaa lisääntyä, jos kielellisin perustein liitetään toisiinsa yhteenkuuluvat sanat, kuten *hoito* ja *hoitaminen.* Ongelmana vain on, että muodollisin perustein tehty palauttaminen saattaa johtaa erimerkityksisten termien yhdistämiseen, kun esimerkiksi leksikaalistuminen on vienyt alunperin läheiset termit erilleen. Leksikaalistuminen on jatkumo: toisissa sanoissa merkitysten erillisyys on helpompi todeta kuin toisissa. Päätöstä ei voida tehdä kieliopin perusteella, vaan se on tällöin lähinnä mielipidekysymys.

Nykyisin yhdyssanojen loppuosia on tiedonhaussa hankala löytää. Kun yhdyssanat pystytään automaattisesti jakamaan osiin, voidaan käänteishakemistoon tallentaa sekä koko yhdyssana että sen osat perusmuodossaan. Vaikka saantia näin voidaan lisätä, vaarana on, että tarkkuus heikentyy. Tiedonhakijaa turhauttaa suuresti, jos hän esimerkiksi *tee*-sanaa etsiessään saa vastaukseksi nekin dokumentit, joissa on esiintynyt sana tie*tee*llinen.

Perusmuotojen vaikutus muistitilan tarpeeseen

Eräs perusmuotohakemistojen käyttöä puoltavia tekijöitä on se, että niiden avulla pystytään säästämään muistitilaa. Jokainen taivutusmuotohan on erilainen merkkijono, joten järjestelmä joutuu varaamaan niille jokaiselle erikseen tilaa hakemistossa.

Jos järjestelmän kapasiteetti on pieni ja aineis-

ton määrä suuri, on tilan säästö tärkeää. Siksi englanninkieliselläkin alueella on kehitetty järjestelmiä, joissa saneet typistetään tiettyjen sääntöjen mukaan. Näin on säästetty noin 25 % järjestelmän muistitilaa (Brzozowski 1983). Runsaammasta pääteaineksesta johtuen vie suomenkielisen tekstin tallennus suhteessa enemmän muistitilaa kuin vastaava määrä englanninkielistä tekstiä. Suomessa muistitilan säästön siis pitäisi olla tuntuvampi.

SITRA on tutkinut tiedonhakujärjestelmän muistitilan käyttöä noin 250 000 sanetta sisältäneellä sanomalehtitekstillä. Sanomalehtitekstissä sanamuotojen palauttaminen perusmuotoon vähensi selvästi hakemistotermien määrää. Perusmuodoissaan tallennettujen hakutermien määrä voi olla 19-53 % pienempi kuin samaisten termien määrä ilman perusmuotoon palautusta, merkkijonomuodossa tallennettuna. Määrät vaihtelivat sen mukaan, mitä sanaluokkia MORFOn analyysiin oli otettu mukaan ja oliko yhdyssanoja ositettu ja osia jälleen kombinoitu.

Järjestelmäarkkitehtuurit

Seuraavassa on hahmoteltu kuusi eri vaihtoehtoa toteuttaa vapaatekstihaun salliva tekstitiedonhallintajärjestelmä.

A. Perinteiset tekstitietojärjestelmät

Tekstin sanat tallennetaan sellaisenaan merkkijonoina käänteistiedostoon. Hakuvaiheessa termit typistetään ja järjestelmä etsii niiden kanssa täsmäävät merkkijonot käänteistiedostosta.

B. Taivutusvartaloiden automaattinen tuottaminen

Järjestelmä on muuten sama kuin edellä, mutta siihen on liitetty ohjelmamoduli, jonka avulla hakija voi halutessaan tuottaa annetun hakutermin perusmuodosta taivutusvartalot.

C. Taivutusvartaloiden tuottaminen ja termien seulominen

Järjestelmä on muuten kuten edellä, mutta lisäksi järjestelmän poimimat käänteistiedoston termit tarkistetaan siten, että hakuvartaloihin täsmäävät käänteishakemiston termit palautetaan perusmuotoonsa. Jos tämä muoto ei ole sama kuin alkuperäinen hakutermi, hakemiston termi hylätään.

D. Perusmuotoiset käänteistiedostot

Saneet palautetaan perusmuotoon ennen kuin järjestelmä muodostaa käänteistiedoston. Hakuvaiheessa käyttäjä antaa sanan perusmuodon ja järjestelmä etsii käänteistiedostosta vastaavan sanan ja edelleen siihen yhdistetyt dokumentit.

E. Perusmuoto- ja taivutusmuotohakemiston yhdistelmä

Jos perusmuotoihin palauttava ohjelma pystyy analysoimaan saneen, se palautetaan perusmuotoon ja näistä perusmuodoista tehdään oma perusmuotohakemistonsa. Jos taas sanetta ei pystytä analysoimaan, se tallennetaan sellaisenaan erilliseen taivutusmuotohakemistoon.

Hakuvaiheessa hakutermejä etsitään ensin perusmuotohakemistosta. Jos niitä ei löydy siitä, varmistetaan, että tiedonhakijan antama hakutermi on ohjelman ymmärtämässä perusmuodossa. Mikäli näin on, ohjelma tuottaa hakutermistä taivutusvartalot ja näitä vartaloita hakutermeinä käyttäen käydään läpi taivutusmuotohakemisto. Järjestelmään voidaan vielä liittää mahdollisuus seuloa löydökset kuten C-vaihtoehdossa.

Ongelmana ovat kirjoitusasultaan samanlaiset eli homografiset sanat, joita on vaikea jäljittää sanastosta — tämä pätee erityisesti yhdyssanoihin. Esimerkiksi väärästä tulkinnasta johtuvaa *keskusohjatimeä*-muotoa ei mitenkään pystytä löytämään hakemistosta *keskusohjain*-sanan eri taivutusvartaloilla.

F. Perusmuoto- ja taivutusmuotohakemiston yhdistelmä, jossa voidaan etsiä väärintulkitut muodot

Tekstin tallennus tapahtuu kuten edellä. Hakuvaiheessa termejä etsitään aluksi perusmuotohakemistosta. Mikäli niitä ei löydy sieltä, varmistetaan, että termi on ohjelman käyttämässä perusmuodossa. Sen jälkeen tulkitaan, onko ohjelma voinut tulkita kyseisen sanan väärin joksikin sen homografiksi.

Homografeja etsitään saman ohjelmiston avulla, joka alunperin teki perusmuotoihin palauttamisen. Prosessi etenee kuten alkuperäinen tunnistusprosessi; siinä etsitään sanastosta sellaisia sanaston sanoja, joiden jotkin taivutusmuodot ovat kirjoitusasultaan samanlaisia kuin hakutermin taivutusmuodot. Mikäli päällekkäisyyksiä löytyy, on näihin sanaston sanoihin tallennusvaiheessa voitu virheellisesti yhdistää jokin hakutermin taivutusmuodoista. (Koskenniemi 1985b, s. 92.) Perusmuotohakemistosta poimitaan nämä mahdolliset väärintulkinnat. Käyttäjän antamasta termistä muodostetaan myös hakuvartalot ja niitä etsitään taivutusmuotohakemistosta.

Lopuksi

Ei ole tarkoituksenmukaista, että tiedonhakijat joutuvat pohtimaan haun kannalta epäolennaisia seikkoja kuten kielen taivutussääntöjä. Tämä ei koske vain suomen kieltä käsitteleviä järjestelmiä vaan pätee muihinkin kieliin. Vaikka jokainen ihminen osaa puhua äidinkieltään, on sen sääntöjen pohtiminen käytännössä hankalaa. Vielä hankalampaa on tiedonhaku vieraalla kielellä, koska sen sääntöjä tunnetaan vähemmän.

On siis järkevää pyrkiä rakentamaan tietojärjestelmiä, jotka eivät vaadi käyttäjältään kovin yksityiskohtaista kielitietoutta. Tämä on sitäkin tärkeämpää, kun ei-ammattilaiset tekevät itse tiedonhakujaan.

Vapaatekstitiedonhaun teoreettisia periaatteita, kuten sitoutumista luonnolliseen kieleen lienee tutkittu sangen vähän. Tämä olisi tarpeen, koska tekstejä tullaan tulevaisuudessa tallentamaan yhä enemmän elektroniseen muotoon (vrt. Ahlbäck 1985). On jopa esitetty, että vapaatekstihaku vaatii enemmänkin tekstin kieliopillisen rakenteen tuntemista kuin aihealueen käsiteanalyysiä (Jackson 1983). Taivutusvartaloita tuottavat ohjelmat voidaan varsin helposti liittää jo olemassaoleviin järjestelmiin. Näin MINTTUunkin voisi saada komennon, jossa hakutermi typistetään automaattisesti. Perusmuotoja tuottavien ohjelmien soveltaminen on käytännössä vaikeampaa.

Jos esimerkiksi ruvetaan palauttamaan tekstin sanoja perusmuotoon ennen käänteistiedoston muodostamista, olisi selvitettävä, miten paljon tekstiä voidaan normalisoida olennaisen tiedon katoamatta. Huomiota olisi kiinnitettävä muun muassa siihen, mitä käyttäjät pitävät termin perusmuotona, koska se ei ole mikään itsestäänselvyys.

Vaikka sananmuotoja analysoivat ohjelmat eivät vielä ole virheettömiä, niitä voidaan kehittää. Automaattisella analyysillä on tällöin se etu, että virheet tehdään systemaattisesti. Eri ihmisten tekemät virheet ovat yhtä monivivahteisia kuin heidän taitonsakin.

Hyväksytty julkaistavaksi 17. 6. 1986

Kirjallisuutta:

Ahlbäck, T. 1985. Indexspråk och tesaurusproblematik. Kirjastotiede ja informatiikka, 4 (2): 53—57.

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

Brzozowski, J. P. MASQUERADE: Searching the full text of abstracts using automatic indexing. Journal of Information Science, 1983 (6): 67—73.

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

Jackson, L. 1983. Searching full-text databases. 7th Int. Online Meet. London, 6—8 Dec. 1983. Oxford, Learned Information. 1983: 419—425.

Jäppinen, H. et al. Kielikone tulkitsee suomea. Helsinki, SITRA, Kielikone-projektin julkaisuja, Sarja C, Selvitykset nro 1. 1985a.

Jäppinen, H. et al. Morphological analysis of Finnish word forms. Selected reprints. Helsinki, SITRA, Publications of the Kielikone-project, Series A, report no. 1. 1985b.

Karlsson, F. Suomen kielen äänne- ja muotorakenne. Porvoo, WSOY. 1983.

Karlsson, F. (toim.) Computational morphosyntax. Helsinki, University of Helsinki, Department of General Linguistics, Publications no. 13. 1985a.

Karlsson, F. Linguistic computer corpora and programs at the University of Helsinki. Helsinki, University of Helsinki, Department of General Linguistics, Publications no. 14. 1985b.

Koskenniemi, K. Two-level morphology: A general computational model for word-form recognition and production. Helsinki, University of Helsinki, Department of General Linguistics, Publications no. 11. 1983.

Koskenniemi, K. An application of the two-level model to Finnish. Teoksessa: Karlsson, F. (toim.). Computational morphosyntax. Helsinki, University of Helsinki, Department of General Linguistics, 1985a: 19-41.

Koskenniemi, K. FINSTEMS: A module for information retrievl. Teoksessa: Karlsson, F. (toim.). Computational morphosyntax. Helsinki, University of Helsinki, Department of General Linguistics, 1985b: 81—92.

Nurminen, Riitta. Suomen kielen sananmuotoja tulkitsevien ohjelmien hyödyntäminen tiedonhakujärjestelmissä. VTT Tutkimuksia 386, Espoo 1986.