

ÄIDINKIELEN OPETTAJAIN LIITON
VUOSIKIRJA XXXIV

Koulu
tietoyhteiskuntaan

Toimittanut
Matti Sinko

ÄIDINKIELEN OPETTAJAIN LIITTO · HELSINKI · 1987

Tietokoneolingvistiikka ja äidinkielen opetus

FRED KARLSSON

Mitä tietokoneolingvistiikka on?

Tietokoneolingvistiikka (engl. computational linguistics, jatkossa lyhennettynä TKL) on melko uusi monitieteinen tutkimussuunta, joka kieliopereen ja tietojenkäsittelyopin menetelmin selvittelee luonnollisten kielten rakennetta ja käyttöä. Kielien rakenteen kuvaukset pyrittään hahmottamaan tietokoneohjelmiksi, jotka päättelivät mm.:

- mitkä ovat tekstissä esiintyneiden sanojen perusmuodot, päätteet ja merkitykset suhteessa kuvauksessa käytettyyn sanakirjaan,
- mikä on tekstin lauseiden rakenne eli jäsenyys,
- miten lauseen merkitys voidaan koostaa funktiona sanojen merkityksistä ja lauseen rakenteesta,
- miten lauseen merkitystuokintaan vaikuttavat muut kuin kielelliset seikat, erityisesti ihmisen niin sanottu arkitieto (engl. common sense knowledge) sekä tieto viestintätilanteesta, sen osanottajista ja aiemmin puheena olleesta,
- mikä on kielellisen tiedon esitysmuoto eli representatio,
- miten kokonaisen tekstin merkitys on ilmaistavissa ja tiivistettävissä,
- miten sanoja ja lauseita voidaan tuottaa merkitysrakenteista ja ihmisten tarkoituksperistä käsin.

Kuten yleisessä kielitieteessä muutenkin, keskeistä on pyrkimys yleisiin ja kattaviin teorioihin. Hyvä yleinen tietokoneolingvistinen teoria on sellainen, jonka asettamien periaatteiden ja formalismien avulla voidaan kuvata moniajaja ehkä (ainakin periaatteessa) kaikkia luonnollisia kielitä. Yhdenmukaisella tavalla eri kielistä tehdyt kuvaukset voidaan sisällyttää yhteen ja samaan tietokoneohjelmaan, joka konkreettisesti suorittaa em. kaltaisia tehtäviä.

Asken mainitut TKL:n keskeisalueet ovat pajjolti päällekkäisiä niiden alueiden kanssa, joita kielitieteessä muutenkin tutkitaan. Ero ilmenee ennen muuta siinä, että TKL:ssa viime kädessä tavoitellaan toimivaa tietokoneohjelmaa. TKL suo uudenlaisia mahdollisuuksia tarkastella kielitä toimintana, tai ainakin simuloida tätä toimintaa tietokoneoh-

kehien avulla. Tällaista simulointien suhde ihmisen kielenkäytön todellisiin prosesseihin on kuitenkin ongelmallinen. TKL ei missään tapauksessa voi väittää korvaavansa esim. empiristä psykologiyksikkä, jonka ensisijaisia tutkimuskohteita nuo "todelliset prosessit" ovat.

TKL on ytimehään teoreettista perustutkimusta, jonka käsitteistö ja menetelmät ovat sekä lingvistiä että tietojenkäsittelyopillisia. Toisaalta TKL:lla on melkoisesti käytännön sovellusmahdollisuuksia: automaattinen kääntäminen (jossa ihmisen osuus aina on melkoinen, viimeistään tarkistusvaiheessa), tekstien tiivistäminen ja indeksointi, in-formaation haku, oikeinkirjoituksen tarkistus, tyylitopin seikkojen ja tekstin luettavuuden arviointi, luonnollisen kielen käyttö tietokantaky-selyjen kielenä, tietokoneavusteinen leksikografia ja kielen opetus ym.

Joskus esiinyy käsitys, että TKL on ensisijaisesti suurten kielitai-neisiojen automaattista muokkaamista ja käsittelyä, esim. aakkosustaa tai ponninaa. Nämä ovat sinänsä hyödyllisiä tietokoneen käyttötapoja. Yleensä niillä ei kuitenkaan pyritä varsinaisten teoreettisten tietokone-lingvistiisten ongelmien valaisemiseen. Tehokas tutkimusaineiston muokkaus on tietenkin arvokas apu kelle tahansa tutkijalle, mutta mikä tahansa kielellisen korpuksen tietokonekäsitely ei silti ole tietokone-lingvistiikkaa. Äidinkielenopetuksessa hyvillä korpuksilla on kuitenkin monenlaisia käytännöllisyyksiä, joita tuonnempaa valote- taan.

Teoreettisen TKL:n tyyppilliset tutkimusongelmat ovat tällaisia: Mi- ten eri kieliooppiformaalismit (kuten dependenssikieliooppi, kategoriakie- lioppi, generatiivinen kieliooppi, erilaiset funktionaaliset kieliooppi) so- veltuvat tietokoneessa implementoitavien eli toteutettavien sääntöjen tai prosessien rungoiksi? Kuinka laskennallisesti tehokkaita no. mallit ovat? Mitkä ovat niiden täsmälliset formaaliset ominaisuudet? Minkä- laisilla jäsenyysratkaisuilla erilaisia kieliooppimaloja ja niiden kehityk- sissä tehtyjä kieliopeja voidaan jäsentää (esim. "top-down" käynnien kielioopin sääntöjä arvauksina löydettyjä rakenteesta, tai "bot- tom-up", jolloin jäsennettävä lause suoremmiin ohjain analyysia)? Miltä- laisella formaalisilla sanojen morfologiser, syntaktiset ja semantiset ominaisuudet pitäisi esittää sanakirjassa? Kuinka pitkälle tietokone- lingvistinen teoria voidaan viedä sellaiseen subntaan, että kielioopin kir- joittaja vapautetaan ohjelmoinnin ja ohjelmointikielen "alempien" ta- sojen lingvistiisesti enemmän tai vähemmän epäolemaisista yksityis- kohdista? Tavoitteena on tällöin, että kielioopin kirjoittaja voisi keskii- tyä ongelman lingvistiiseen puoleen ja kerran tehty ohjelma hoitaisi au- tomaattisesti mahdollisimman suuren osan yksityiskohdista.

TKL on viime vuosina kasvanut nopeasti monissa maissa. Suo-

nessa laajempi tutkimus käynnistyi vuoden 1981 alusta, jolloin do- senti Kimmo Koskeniemi ja minä aloitimme yhteistyön Suomen Akatemian rahoituksen tutkimussopimuksen vahvistamana. Aluksi työmme kohdistui ensisijaisesti morfologiaan (ks. tarkemmin Kosken- iemi (1983) ja kokoomateoksen Karlsson (1985, toim.) artikkeliei- län). Sittemmin perustettiin Suomen Akatemian rahoituksella Helsin- gin yliopiston tietokone-lingvistiikan tutkimusyksikkö viisivuotiskau- deksi 1985-1989. Tutkimusyksikkö tähtää ennen kaikkea yleispatē- vän tietokone-lingvistiisen syntaktisen analyysiformaalismin kehittämi- seen. Lopullisen mallin tulee olla morfologiaan integroitu, avoin se- mantikan suunaan sekä pysyvä käsittelemään normaalia juoksevaa tekstiä. Yksikköön kuuluu kuusi työntekijää ja se toimii lähesssä yh- teistyössä Helsingin yliopiston yleisen kielitieteen laitoksen ja usei- den ulkomaisen laitosten kanssa.

Yleisiä diidaktisia näkökohtia

Tietokoneet ovat apuneuvoja, työkaluja, käytöitään tosin hankalam- min hallittavia kuin tavanomaiset työkalut. Työkalan pitää tehostaa, nopeuttaa ja muutenkin edistää varsinaisen tehtävän suorittamista, muussa tapauksessa sen käyttö tuskin on mielekästä. Opetuksen kan- nalta täytyy lisäksi vaatia, ettei työkalu orjuuta oppilaita tai opettajia drillinomatilla. Luovuuden tappavilla, ennalta jökseenkin tarkkaan määrättyillä stereotyypisillä tehtäväsarjoilla. Varoitavan esimerkin tarjoavat vieraiden kielen opetuksen 1960- ja 1970-luvulla rakenne- tutkielstudiot. Alkuvaiheessa uskottiin yleisesti, että kielistudiot mut- listaisivat vieraiden kielen opetuksen ja ratkaisevasi parantaisivat op- pimisuloksia. Pian kuitenkin osoittautui, että varsinaisen kompatus- kivi on tarpeeksi monimuotoisten, vaihtelevien, luovuutta vaativien ohjelmien ohjelmien aikaansaaminen. Aivan hian usein kielistudio- ohjelmat ovat vieläkin mekaanisia morfologis-painotteisia drilliharjoi- tuksia tai yksinkertaisia syntaktisia transformatioharjoituksia (myön- tölause → kielitölause, aktiivilause → passiivilause, väitelause → ky- symyslause jne.).

Ennen kuin siirryn käsittelemään varsinaisen TKL:n tarjoamia ope- tussovellusmahdollisuuksia, mainitsen sen itsestään selvyuden, että tavallinen tekstinkäsittely on kaikkein ilmeisin tapa aloittaa tietokonei- den käyttö äidinkielen opetuksessa. Tekstinkäsittelyohjelmia ja niiden apuohjelmia, tai jopa käyttöjärjestelmän komentoja, voi hyödyntää paitsi kirjoitelmien jatkuvan paranteluun myös tiettyjen sanojen ja il- mausten etsimiseen tekstistä (esim. virheiden metsästämisessä), indek- sien laadintaan, aakkosutukseen ja frekvenssien laskemiseen.

Tekstinkäsittelyä voi rydyttää muunkinlaisilla ohjelmilla. Kaupalli-

sesi saatavana on jo SITTRAN Kielikone-projektin valmistama suomenkielisen tekstin oikelukujärjestelmä TekoPLUS, joka edellyttää kovalevyn käyttömahdollisuutta. Lähtöleveysuudessa saatavana ovat kuitenkin Tietokone-lingvistiikan tutkimuskesikön tekemä lähelykeko-neisinkin mahuvat suomen kielen oikelukujärjestelmä. Samaten kohta saatavana on suomen kieltä varten ohjelma, joka tarkistaa mm. vierasperäisten sanojen oikeinkirjoituksen (prototyypit on jo valmistettu). Tietokone-lingvistiikan tutkimuskesikön tekemänä) sekä ohjelma, joka tarkistaa vyyliopin kekkiset seikat. Vuoden tai kahden kuluttua suomenkielisen tekstin käsittely-ajapuvuoina on siis melkoinen joukko hyötyohjelmia. (Laadukas suomen kielen automaattinen tavutus on ollut olemassa jo useita vuosia (Karlsson 1985b).)

Ketvöllisten kieliohjelmiin tekeminen äidinkielenopetuksen tarpeisiin edellyttää paitsi tietokone-lingvistiistä myös pedagogista ja didaktista asiantuntemusta. Jatkossa esitelen lähinnä mitä TKL:n näkökulmasta nyt on tarjolla ja mitä realistisesti voitaisiin sisällyttää suomen kielä käsittelyyn opetusohjelmiin. On syytä korostaa, että varsinainen didaktinen ideointi ja toteutus on tekemättä.

Keskeisimpänä näen itse mahdollisuuden luovasti hyödyntää olemassa olevia laajoja korpuksia ja suomen kielen analysohjelmiä. Tämä voisi tuoda jotain uutta äidinkielen opetukseen, nimittäin dynaamisen ja nopean mahdollisuuden testata kieliohjeiden sääntöjä ja rakentaa suurla aitojen tekstien määrillä sekä kehittää oppilaiden kielen rakenteen ja sen elementtien operatiivista tunteesta. Kieliohjeiden omaksuminen ei jäisi vain kirjojen sääntöjen ja niiden usein kankaisten esimerkkilauseiden ulkoistaiseen muistamiseen. Kuiten viimeisessä jaksossa osoitan, on myös aivan realistista antaa oppilaiden itse kehittää pieniä kieliohjeiden katkelmia ja testata niiden toimivuutta.

Tässä kirjoituksessa en ninkään käsittele mahdollisuuksia laatia varsinaisia ohjeiden tapaisia opetusohjelmia, joissa edetään ennalla määrättyillä tavalla tehtävästä toiseen sen mukaan, miten oppilas pystyy tehtävät ratkaisemaan. Näinhän ohjelmoitu opetus on usein ymmärretty. Tämä näkemys on kuitenkin turhan suppea.

Korpuksien käyttö

Tietokoneen luettavassa muodossa on jo saatavana useita sanaluettelotia ja sanakirjoja (vrt. Karlsson 1985a): Suomen kielen käänteis-sanakirja (jossa on Nykky-suomen sanakirjan sanavarat), Suomen kielen taajuussanasto, Slangisanasto, Uudissanasto 80, Sivistyssanakirja ym. Tällaisia korpuksia voi käyttää helposti ja kohtalaisen monipuolisesti yksinkertaisten käyttöjärjestelmän komentojen ja ennalla tehtyjen pikku ohjelmien avulla tarkoituksena erityisesti löytää esimerkkejä,

joita valaisevat esille tulleita lingvistisiä ongelmia tai esitettyjä kysymyksiä. Rajoitus on toki se, että tällaiset haat voivat kohdistua vain fonologisiin ja morfologisiin seikkoihin. Helppo on kuitenkin saada vastaus tällaisiin kysymyksiin:

- Mitkä ovat kielen yleisimmät johdimet, yhdyssanat ja taiputusrypyt Taajuussanakirjan aineiston valossa?
- Mitkä taiputusrypyt esiintyvät erityisen runsastukuisina Taajuussanakirjan yleisimpien sanojen joukossa (eli miten morfologinen kompleksisuus pyrki kasaantumaan yleisimpiin sanoihin?)
- Mitkä sanaluokat esiintyvät erityisen usein yleisimpien sanojen joukossa?
- Mitkä perussanat ovat erityisen suosittuja yhdyssanojen osina?
- Millaisia fonotaktisia rajoituksia eri taiputusrypyissä on, esim. tiettyjen konsonanttiyhymien puuttuminen vartalonloppuisten vokaalin edessä?
- Mitä erityisiä fonotaktisia keinoja (esim. sanansisäisiä ja -alkuisia konsonanttiyhymiä) käytetään slangisanoinssa?

Tällaiset haat voivat monin tavoin edistää oppilaiden kielen rakenteiden, sääntöjen ja kehitystendenssien tajuua. Suuri etu on se, ettei kysymyksiä ole enkkäteen rajattu. Pienellä lingvistisellä alkuopauksella voidaan mielenkiinto ja mielikuviutus ohjata oikeaan suuntaan. Parhaassa tapauksessa voidaan tuottaa uuttakin tietoa. - Edellytyksenä on toki, että opetajalla itsellään on tarvittava perustäkemys sopivien ongelmien viritämiseksi.

Saatavana on myös monenlaisia juoksevaa tekstiä sisältäviä korpuksia. Niistä voidaan etsiä mitä tahansa merkkiomnomotoisesti ilmaistavia esimerkkejä, niin sanojen osia, yksittäisiä sanoja kuin syntagmojakin. On myös mahdollista käyttää ohjelmia, joilla voidaan ilmaista kielen rakenteita eikä pelkästään merkkiomnomia (esim. BE-TA-ohjelma, Karlsson 1985d). Esimerkkien poiminta on siis jo nykyisellään mahdollista hyvinkin sofsitikonomeia tarkoituksia varten.

Vielä mielenkiintoisempia poimintamahdollisuuksia tarjoaa koodattujen korpuksien käyttö. Tällainen on ns. HKV-korpus (Häkkinen, Karlsson ja Vilkuu 1980). Nämä kirjoittajat tekivät seikkaperäisen syntaktisen sekä osittain semanttisen ja tekstuaalisenkin analyysin noin 120 asakirjatekstistä (yhteensä noin 10150 yksinkertaisista lauseista). Joka lauseesta analysoitiin 63 muuttujaa. Muuttujat ovat keskeisten nominaalijäsenten (subjektin, objektin, predikaatiivin) kohdalla mm.: konstituenttirakenne (yksi sana, etuattribuutti + pääsana, useita etuattribuutteja + pääsana, nominaalisuus, lause, infinitiivi jne.), sija (nominatiivi, partitiivi, genetiivi), tunnetuus (mainittu, implikoitu, edellytetty kielenulkoisen tiedon perusteella ym.), tekstuaalinen sidos-

keino (ellipsi, anafora, katafora, toisto, välikäsite ym.) ja sanamäärä. Verbyhymästä on erotettu mm. yksisanaiset (vain finitiiviverbi), liitomuodot, infinitiivirakenteet ja tiionnit. Kaikista lauseista on analysoitu mm. lausetyyppi (päälause, et sivulauseyyppi), lauseen modus (väite, kysymys, kiisky, haudhhdus), generisyys, pragmatiset elementit, polaarisuus (myöntö, kieltö), lauserakenne (mm. transitiivisuuden ja päähuokan yhdistelmät, omistusrakenteet, kausatiivirakenteet), konstituenttien ja sanojen määrät, adverbiaalien määrä sekä pin-täjäisyys SVO-tyyppisten kaavojen avulla.

Välittömän hakuohjelman avulla korpuksesta voi tuostaa ne lauseet, joissa haluttu muuttujan yhdistelmä toteutuu. Esimerkkejä mahdollisista poiminnoin:

- Passiivilauseet, joissa objekti on lauseen alussa.
 - Eksistentiaalilauseet, joissa partitiivimuotoinen subjekti on lauseen alussa.
 - Lauseet, joissa on vähintään neljä adverbiaalia.
 - Transitiivilauseet, joissa objekti on lauseen alussa ja subjekti edeltää verbiä.
 - Lauseet, joissa tunnettuudeltaan uusi subjekti esiintyy lauseen alussa.
 - Predikaatiivilauseet, joissa predikaatiivina on N + lause.
 - Lauseet, joista verbi elliptisesti puuttuu.
 - Lauseet, joissa objekti on topikaalisitettu emfaasin takia.
 - Lauseet, joissa subjekti ja finitiiviverbi eivät kongruoi.
 - Transitiiviset kieltolauseet, joissa objektin sija on muu kuin partitiivi.
 - Kieleteiskysyviä lauseet, joissa kysymys kohdistuu adverbiaaliin.
- Kuolen näkyä, hakuja voi tehdä hyvin monenlaisia. Tällaisten aineistojen avulla äidinkielen opetuksen kielitopin osiota voidaan huomattavasti elävöittää. On mahdollista esittää kaikenlaisia kysymyksiä mahdollisista lauseista ja kielen sallimista rakenteista ja siten korpuk-sen avulla verifioida, esiintyykö tällaisia rakenteita vai ei. Tällä tavoin kieltä problematisoimalla voidaan parhaassa tapauksessa saada hyvin-kin luova ja eksploraatiivinen ote opetukseen.
- Edellytyksenä HKV-korpuksen käytölle on kovalevykone. Muutama vuoden päästä näitä alkaa varmasti ilmestyä kouluihinkin. Silloin edellä luonnosteltu laajan koodaan korpuksen käyttö ei ole utopiaa.

Morfologia

Muoto-oppi on äärellinen ja selväpiirteinen ja siten suhteellisen helposti formalisoitava kielen osa-alue. Tietokoneingvitysisen morfologian kiintoisin tulos tältä vuosikymmeneltä on Kimmo Koskenniemen

väitöskirjassaan (1983) esittämä ns. morfologian kaksisanomainen (Ks-myyös Koskenniemi 1985a,b). Tämä on ainoa toistaiseksi tunnettu menetelmä, jonka avulla voidaan miltei tahansa kielelle laadita toimivia sanonjen laipunnisen kuvauksia. Näin tehdyt kuvaukset voidaan siirtää tietokoneohjelmaan, joka sekä tunnistaa että tuottaa asianomaisen kielen sanamuotoja. Tunnistaminen tarkoittaa erityisesti sitä, että määrätään sanamuodon perusmuoto ja identifioidaan kielipöhlisin määrätään kaikki sen sisältämät päätteet. Esimerkkejä suomen kielen sanojen tunnistamisesta (morfologisesti monitulkintaisista sanoista löytyvät kaikki tulkinnat):

SANA	MORFOLOGINEN TULKINTA	
kolojen	kolo	SUBST MON GEN
rikoampi	rikos	ADJ COMP YKS NOM
rikoin	rikos	ADJ POS MON INSTR
	rikos	ADJ SUP YKS NOM
tehtävä	tehtävä	SUBST YKS NOM
	teke	VERBI PSS PCP1 YKS NOM
suoritto	suoritto	VERBI AKT PRES IND YKS 3 PERS
	suoritto	VERBI INF 1 NOM
olusto	olusto	SUBST YKS NOM
	olus	SUBST YKS NOM
	olunen	SUBST YKS PT
	olku	SUBST YKS ELA
	olusto	VERBI IMPV YKS 2 PERS
	olusto	VERBI NEG
loinhuuto	loin	SUBST YKS GEN + huuto SUBST YKS NOM

Tietokoneen luettavassa suomen kielen kuvauksen sanakirjassa on tällä hetkellä noin 15.000 sanaa. Kaikkien näiden sanojen kaikki muodot voidaan tunnistaa ja tuottaa. Kuvaus kattaa myös johto-opin ja yhdysanat. Kuvauksen edellyttämää sanakirjaa voidaan kätevästi kasvattaa lisäämällä uusia sanoja. (Sanamuotoja voidaan tunnistaa vain edellyttäen, että ne on sisällytetty kulloinkin käytössä olevaan sanakirjaan, joka on yksinkertainen luettelo sanojen perusmuotoja mahdollisista morfologisista erikoismerkinnöineen.) Kaksisanomainen 10.000 sanan sanakirjoin toimii jo 360 kB:n levykoneissa.

Niiden avulla olisi mahdollista toteuttaa monenlaisia morfologisia opetussovelluksia. Oppilaan tehtävänä voisi olla esimerkiksi:

- nimetä annettujen tai hänen itsensä keksimien sanamuotojen sisältämät päätteet (esimerkkejä edellä),
- tuottaa annetuista sanoista ja kielipöhlisistä määrätellyistä päättesar-

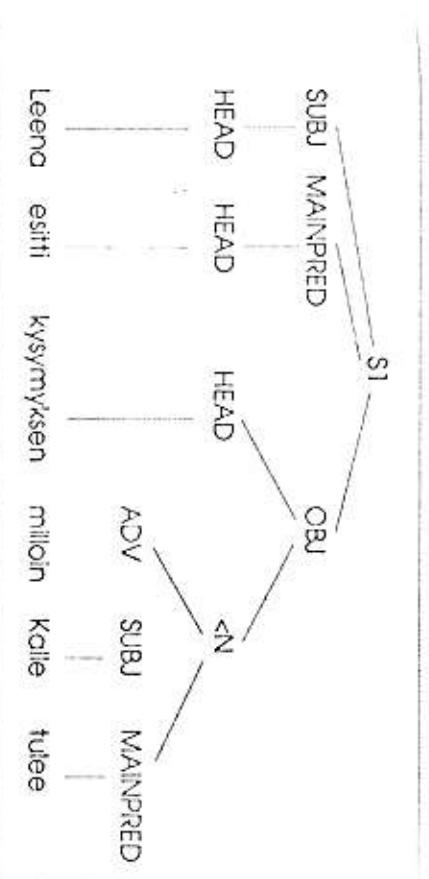
- joista vastaavat konkreettiset sanamuodot, esim. "rikas MION PTIV" = rikkaana, "anta PSS PCP2 MION PTIV" = annettuina.
- tunnistaa annetuista tekstinistä kaikki tietyn sanaluokan tai muun morfologisen luokan sanat;
 - tunnistaa annetuista tekstinistä kaikki sen yhdyssanat ja analysoida ne osiinsa;
 - tunnistaa annetuista tekstinistä mahdollisimman paljon morfologisesti monitulkintaisia sanoja;
 - tunnistaa annetuista tekstinistä kaikki tietyntyyppiset johdokset;
 - pohdita, miksi annettujen tekstien jokin sana ei välttämättä kuulu analysoituihin (nimitettiin koska ne eivät satu sisällymisiin käytössä olevaan sanakirjaan, koska ovat harvinaisia ammatitermejä, arkaisimpia, uudissanoja tms.);
 - tarkistaa, onko tietty muoto jonkin sanan norminmukainen muoto, esim. valtioa, rikkahin, suklaaseen, möin,
 - päätellä, miten uusi sana (esim. laina- tai slangisana) produktiivisten taivutusääntöjen mukaan taipuisi,
 - tuottaa annetuista vartaloista tietyntyyppiset johdokset.
- Luettelo ei suinkaan ole tyhjentävä. Joka tapauksessa ohjelma arvioisi oppilaan suoriutuksen oikeellisuuden ja tason sekä antaisi sopivaa palautetta. Tämänkaltaisia opetusohjelmia on varmasti saatavana viimeistään vuoden 1987 loppupuolella (sen sukunimi on jo valmistunut AOL:n SJAINEN, joka harjoituttaa sijoja).

Syntoksi

Syntaksin täydellinen, tai edes tyydyttävä, tietokoneplementointi on hauskampaa kuin morfologian. Syyt ovat ilmeiset: syntaksi ei ole sillä tavoin äärellinen ja selkeäpiirteinen kielen osa-alue kuin morfologia. On vaikea täsmällisesti sanoa, mikä kielen (kaikki) syntaktiset säännöt ovat. Erityisen hauskalla tavanomaisen kielen kuvauksen tasolla ovat sellaiset ilmiöt kuin ellipsis, erikoiset sanajärjestykset (esim. topikalistuksesta johtuva), etäiset viittaukset (kun pääsanat ja määritteet eivät olekaan vierekkään syntagmassa), infinitiivirakenteet ja lauseenasitukset sekä runsaasti sivulauseita sisältävät konstruktiot.

Täistä huolimatta jokseenkin täydellinen suomen kielen automaattinen lauseenjäsennin eli parseri on jo olemassa (Karsson 1985c). Jäsennin toimii laajan (noin 10.000 sanaa sisältävän) sanakirjan avulla ja pysyy tunnistamaan kaikkien näiden sanojen kaikki muodot. Morfologisia monitulkintaisuuksia jäsennin pystyy monessa tapauksessa ratkomaan. Syötetyille lauseille kehitetään funktionaalinen jäsennyspuu, joka paljolti muistuttaa klassista lauseenjäsennystä. Joka lausees-

siä tunnistetaan siten subjekti, objekti, predikatiivi ja adverbiaali. Nämä kuvaukset annetaan pääsanalle. Edelleen joka pääsanalle tunnistetaan sen määritteet. Myös infinitiivirakenteet ja lauseenasitukset pystytään oikein tulkitsemaan, summaen toisessa lauseessa tai nominaalisesä konstituutissa jäsenenä oleva lause. Lause Leena esitti kysymyksen, milloin Kalle tulee saa seuraavan kuvauksen (MAINPRED = ti- näittäverbi, HEAD = pääsanana, <N vasemmalla olevan substantiivin attribuutti):



Tämä jäsennin toimii toistaiseksi vain isoissa ns. Lisp-tietokoneissa ja kysymyksessä on ensimmäinen prototyyppi. Vaativat mitat täyttävä, mikro tietokoneessa toimiva automaattinen lauseenjäsennin ei ole aivan ovelta, mutta muutaman vuoden kuluttua sellainen lienee jo rakennettu. Siinä vaiheessa avoimet ja luovat opetussovelhuksetkin käytävät realistisemmiksi. "Avoimella" ja "luovalta" tarkoitan nimenomaan sitä, ettei kysymyksessä ole pelkästään ennalla annettujen tai rakenteeltaan kovin rajoitettujen lausien analysointi. Sellaisia syntaktisia ohjelmia on toki mahdollista tehdä jo nyt.

Esimerkki: valintorajoitukset ja kielipöpin teko

Lopuksi esittelen tekemääni ohjelmaa, jonka tarkoituksena on kirvoittaa käyttäjää pohdintaan, minkälaisia semanttisia rajoituksia lauseiden täytyy noudattaa, ja joka samalla suo mahdollisuuden itse laatia yksinkertaisia kieliohjeita ja testata niitä. Ohjelma tuottaa satunnaisesti muodostettuja suomen kielen pintasyntaksin mukaisia "lauseita", jotka kuitenkin sisältävät runsaasti rikkomuksia sanojen välisiä valinta- eli se-

lektioirjoituksia vastaan. Tässä satunnainen valikoima "lauseita":

Viivo si on sorkka on a ng t a r v n u t v a r s n a s e n v i n e n .
 H a i n n e e p i n m i e l e t a k o p o s t a t o o n t u n n u s m u n s i m a l d o n .
 M a r i n o k e l j a p u o l i t e n e e h o u k a t e t u r j a f r e e .
 K a s a n h e v o i s o p p i n e e n s a k a i v a i t t o m a s e t t i t v o r o i l l i s u s s e r o .
 H e m e k e t o t o r j o s t u o n v i e r a n f o r m o m i n .
 T u o e x l e t i n e n m a k a r a s e n o i s e n .
 P o n k e e n y t i s e v a i t e t a k e n t a a v i n h e e n .
 M a r j a n e n E n o o n p u h t o o k s i p e s t y .
 S e n t i o n A n t t i o n o p p e s i n i .
 O j a l a u s u k a r v a h u u m a o v e n h e l p o n G u n n a r i n .
 S i e l u j o k a o n p o j a s k i e m u r o o n m i e l i .
 I l o n e n j a T a i n e n k a u n o s i e l u o n t o s k u n p o r j a .
 M u n a m a k a o n k a u r o p u u r o o n k e r t a k a k k e n R o m b o .
 K v i o t o l l u i m u t t a i n m i t t a n e n d e m o k r a t i a .
 L a t t o n h u n e n o m e n o m e n t o n t y k i v e e n .

Tällaiset "lauseet" ovat syntaktisesti niin hyvämuotoisia, ettei niiden rakenteen tunnistaminen yleensä tuota vaikeuksia. Semanttiikkaan sen sijaan ei normaalisti ole täyttä tulkua. Toisaalta syntagmaattiset semantitset rikkeet ovat niin selkeästi havaittavissa ja paikallistettavissa, että niiden pohdiskeleminen ja todellisten rajoitusten formulointi on mielekästä. Tällaisesta aineistosta on helppo todeta, mikä rajoitukset ja tulkintaprosessit pätevät normaalikielenkäytössä. Tietyt verbit saattavat edellyttää subjektinsa ja/tai objektinsa tarkoitteita esim. elollisuutta, ihmillisyyttä, agenttiivisuutta, mentaalisuutta tai konkreettisuutta; omistuneesti käytetyt adjektiivit edellyttävät usein substantiivipääsanoihaan määräämätönmuksia; vertailu edellyttää vertailtavien yhteismatalisuutta; kontradiktiot (kuten "mutanen – puhtaksi pesty") tuntuvat yleensä anomaahsilta; eräissä tapauksissa ihmiselle tyyppillinen analoginen ja metaforinen ajattelu saattaa tuottaa hyväksyttäviä tulkinnoja, hyvinkin yllättäville yhdistelmille jne.

Tällätavoin voidaan melko luovasti ja hauskasitkin havainnollistaa vaikeita kieleen liittyviä rajoituksia ja prosesseja ja motiivoida oppilaita pohdiskelemaan mistä rikkeet johtuvat. Yhdellä ainoalla käskyllä voidaan haluttaessa tuottaa toivottu määrä lisäesimerkkejä.

Askel vielä vaativampaan suuntaan olisi antaa oppilaiden itse rakentaa kieliohjelma, joka tuottaa lauseita. Tämä on helpompaa kuin kuvitteluksi. Po. ohjelma tulkiisee ns. kontekstivapaita (context-free eli CF) sääntöjä. Käyttäjän ei tarvitse välittää ohjelmoimista, riittää kun hän formuloi haluamiaan sääntöjä sovitun formalismin mukaan. CF-sääntö tarkoittaa yksinkertaisesti, että yksi symboli toisinkirjoitetaan yhdeksi tai useammaksi toiseksi symboliksi (jotka voivat olla joko toisia toisinkirjoitettavia symboleja taikka lopullisia sanoja). Seuraava yksinkertainen CF-kieliohjelma tuottaa osajoukon suomen kielen trans-

tiivilauseita <kuuhmasuikheet tarkoitavat, että valitaan yksi ja vain yksi elementti jo. joukosta>. Ensimmäinen symboli toisinkirjoitetaan aina oikealla olevaksi.

S N P r o m s g V P
 S A P r o m s g V P A D V
 N P r o m s g N r o m s g
 N P r o m s g A n o m s g N n o m s g
 N e n s g N e n s g
 N p r o v N d i v
 N e n o m p l N n o m p l
 V P V N r g e n s g
 V P V N P p r v
 V P V N n o m p l
 N r o m s g < n o i n e n a u t o k e n v i k v i k a n s a k a u n o s e u r a k k a u s >
 N e n s g < t a d o n k a r v n m a k k a r o n m i e n e n m i e l e n s >
 N d v < k o n e t t a s e r i o d i o t o o s i a h a y n e n a >
 N o m p l < o p p e t k a p i t a e t i k s e t v i l l i s a t >
 A n o m s g < s r n e n v s a s p % k a k a r v a n e n n e i s k u m o i n e n s >
 V < e n t a a t i l a d a h a u d a t o v o i e s i t t ä k a i s o s i o o s >
 A D V < e l l e n s i i o n j o e r k a k a i k o t o n a s i n n e >

Jos samaa symbolia varten on monta toisinkirjoitusääntöä (kuten esimerkissä V P:itä varten kolme), ohjelma valitsee aina satunnaisesti yhden. Vaikka morfologisia rajoituksia ei tällaisella kovin yksinkertaisella CF-formalismilla olekaan mahdollista kätevästi ilmaista, voidaan silti helposti esittää laajojakin kielitopin kateleimia. Esim. subjektin etumääritteettömään pääsaanahan liittyvärelatiivilause on kuvattavissa tällä rekursiivisella säännöllä:

N P r o m s g N n o m s g S

Olemasta tässä ei ole se, voidaanko lausia täydellinen kieliohjelma, vaan paremminkin se, että oppilaat oman mielikuvituksensa avulla voisivat kokeilemalla päästä selville kielen rakenteesta. Jo sellaiset peruskäsitteet kuin sanaluokat ja lauseenjäsennet konkreettisuudet ja saavastiltoit aivan toisella tavalla kuin päähän päättämällä, jos havaitaan, että ne ovat *rakennusaineita*, josta kieli koostuu ja joita ihmiset käyttävät lauseita muodostaessaan.

Kirjallisuutta

- Hakulinen, Auli & Karlsson, Fred & Viikuna, Maria 1980. Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus. Helsingin yliopiston yleisen kielitieteen laitoksen julkaisuja No. 6.
 Karlsson, Fred 1985. Ioinn. Computational Morphosyntax. Report on Research 1981-1984. Helsingin yliopiston yleisen kielitieteen laitoksen julkaisuja No. 13.

- Karlsso, Fred 1985a. Linguistic Computer Corpora and Programs at the University of Helsinki. Helsingin yliopiston yleisen kieliteleen laitoksen julkaisuja NO. 14.
- 1985b. "Automatic Hyphenation of Finnish". Ks. Karlsso (1985, toim.), s. 93-113.
 - 1985c. "Parsing Finnish in terms of Process Grammar". Ks. Karlsso (1985, toim.), s. 137-176.
 - 1985d. BETA-järjestelmä. Helsingin yliopiston yleisen kieliteleen laitoksen opintotomisteet No. 2.
- Koskeniemi, Kimmo 1983. Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. Helsingin yliopiston yleisen kieliteleen laitoksen julkaisuja No. 11.
- 1985a. "A General Two-level Computational Model for Word-Form Recognition and Production". Ks. Karlsso (1985, toim.), s. 1-18.
 - 1985b. "An Application of the Two-level Model to Finnish". Ks. Karlsso (1985, toim.), s. 19-41.
 - 1985c. "A System for Generating Finnish Inflected Word-Forms". Ks. Karlsso (1985, toim.), s. 63-79.
 - 1985d. "FINSTEMS: A Module for Information Retrieval". Ks. Karlsso (1985, toim.), s. 81-92.