

Fred Karlsson ja
Kimmo Koskenniemi

Ihmisen luonnollinen kieli on hyvä viestintäväline. Se on kenen tahansa käytettävissä, ja sillä voidaan ilmaista mitä asioita tahansa.

Automaattisen tietojenkäsittelyn nopea leviäminen pakottaa selvittämään, miten laajalti luonnollista kieltä, esimerkiksi suomea, voidaan käyttää tietokoneiden kanssa toimittaessa. Näin on tehtävä, jos halutaan, että useimmat ihmiset voivat hallita tietoyhteiskunnan tärkeintä apuvälinettä, tietokonetta.

Koneen kieleksi äidinkieli

Tietokoneiden komento- ja ohjelmointikielien on kehitetty lähinnä siksi, että ihmisen oma kieli sisältää paljon selästä, mitä tietokoneen on vaikea ymmärtää ja tulkita. Luonnollisen kielen ilmaukset voivat olla epätarkkoja ja monimerkityksisiä, eikä kaikkea olennaista aina sanota julki. Luonnollisen kielen merkityksen tulkinta riippuu tilanteesta ja viestintään osallistuvien tiedoista.

Tietokonekielet ovat ihmiselle vieraita

Tietokonekielten tavoitteena on eliminoida kaikinainen epämääräisyys. Niiden avulla voidaan koneelle antaa käskyjä tiiviisti ja täsmällisesti, jotta kone tietäisi yksiselitteisesti, mitä sen pitää tehdä.

Tietokonekielten täsmällisyys on samalla niiden heikkous. Keinotekoiset, formaalit kielet ovat ihmisluonnolle vieraita. Niiden oppiminen ja käyttäminen vaatii ihmiseltä melkoisia ponnisteluja, olipa sitten kyse varsinaisista ohjelmointikielistä tai jonkin tietokonepohjaisen rekisterin tietojensyöttökielestä.

Yksi tietokonekieli kelpaa yleensä vain tietyntyyppisiin sovelluksiin. Tietokoneiden käytön lisääntyessä ihmiset joutuvat tekemisiin yhä useammanlaisien sovellusten kanssa. Erityisen hankalaa on käyttää monia toisistaan poikkeavia tekokieliä. Toisaalta ei voida olettaa, että useimmat ihmiset koskaan oppisivat riittävän hyvin useita tietokonekieliä. Tulevaisuudessa erilaisia tietokoneella suoritettavia tehtäviä on kuitenkin paljon, ja monet niistä ovat mutkikkaita. Luonnollinen kieli näyttää olevan ainoa yleispätevä viestintäkeino ihmisen ja koneen kesken.

Kielentutkijat asialla

Kielen osaamisella tarkoitetaan monia asioita. Osaaminen kattaa kuullun ja luetun ymmärtämisen sekä puheen ja kirjoituksen tuottamisen. Käsittelemme tässä artikkelissa lähinnä kirjoitetun kielen ymmärtämistä ja sen toimintamallien tekemistä eli kirjoitetun kielen mallintamista. Puutumme jossakin määrin myös kirjoitetun kielen tuottamiseen. Onhan ihmisen ja tietokoneen viestintä kaksisuuntaista: ihminen kirjoittaa sanomia koneelle, joka kykyjensä —

ihmisten laatimien ohjelmien — mukaan vastaa.

Puhutun kielen ymmärtämisen teoria ja mallintaminen on hankalampaa kuin kirjoitetun kielen. Se edellyttää mm. äänisignaalin purkamista. Tämä on laaja ja vaikea tutkimuskohde ennen kaikkea siksi, että signaalit ovat melko erilaisia eri ihmisillä ja vaihtelevat muutenkin.

Helsingin yliopiston yleisen kielitieteen laitoksessa on vuodesta 1981 tutkittu suomen kielen mallintamista. Suomen Akatemian rahoittamassa hankkeessa päätavoitteena on luoda kielitieteellisesti perusteltu, kattava ja hyvin toimiva tietokonemalli suomen kielen muoto- ja lauserakenteelle. Mallin täytyy pystyä käsittelemään tavallista tekstiä, jonka sanastoa ei ole ennalta rajoitettu. Mallin pitää pystyä sekä tunnistamaan muotoja — identifioimaan sanojen perusmuodot ja päätteet — että tuottamaan kaikkien sanojen kaikki eri muodot.

Kielen ymmärtämisen tasot

Tietokoneiden markkinoijat saattavat puhua ymmärtämisestä hyvinkin alkeellisten toimintojen yhteydessä — kun laite esimerkiksi tottelee muutamia sovituita englannin- tai suomenkielisiä käskyjä, kuten komentoa "Change A to B", tai kun se pyytää käyttäjää vastaamaan yksinkertaiseen kysymykseen "Tiedoston nimi?".

Tässä ei kuitenkaan ole kysymys varsinaisesta ymmärtämisestä. Kone pystyy tunnistamaan ja tuottamaan vain tiettyjä merkkijonoja niiden täsmällisesti sovitun ulkoisen hahmon perusteella.

Luonnollisen kielen ymmärtäminen on monikerroksinen ja monivaiheinen tapahtuma. Ensinnäkin on oltava selvillä kielen rakenteesta. Tällä alimmalla ymmärtämisen tasolla ihmisen tai laitteen on pystyttävä ainakin seuraaviin ymmärtämisen vaiheisiin:

- Hahmottamaan juoksevan tekstin sananmuotojen sisäinen rakenne (morfologinen hahmotus). Esimerkiksi sana "kädessä" koostuu vartalosta "käde" ja päätteistä "ssä" ja "än".
- Suhteuttamaan eri vartalot ja päätteet kielen käyttäjän sanavaraston, "sisäisen sanakirjan", perusmuotoon (leksikaalinen haku). Esimerkiksi "käde" on perussanan "käsi" yk-

si vartalo, ja päätte "ssä" on sama kuin "ssa".

● Hahmottamaan lauseen rakenne eli siinä olevien sanojen keskinäiset suhteet (syntaktinen hahmotus). Esimerkiksi lause "Tytöllä oli reppu selässä" on melko vakiomuotoinen rakenne, jonka jäsenet ovat adverbiaali "tytöllä", verbi "oli", subjekti "reppu" ja adverbiaali "selässä".

Ihmisaivot purkavat sananmuotojen rakenteen tavalla tai toisella, sillä ei voida olettaa, että kaikki sananmuodot olisivat valmistavaran kielenkäyttäjän sisäisessä sanakirjassa. Muutenhan ei pystyttäisi selittämään esimerkiksi sitä, miksi ihminen yleensä heti pystyy käyttämään oppimiaan uusia sanoja, kuten "kloonit" tai "kvarkki", kaikissa tarvittavissa muodoissaan. Sama pätee myös lauserakenteeseen.

Kirjaimellinen ymmärtäminen

Nämä sinänsä mutkikkaat ymmärtämisen vaiheet, morfologinen hahmotus, leksikaalinen haku ja syntaktinen hahmotus, luovat vasta rakenneluurangon, jonka avulla on mahdollista tulkita tärkein eli lauseen merkitys. Jos järjestelmä, oli se sitten ihminen tai kone, pystyy tunnistamaan sanojen ja lausekokonaisuuksien kirjaimellisen merkityksen, voidaan jo puhua perusymmärryksestä.

Yksityisten sanojen kirjaimellinen merkitys selviää toisaalta sisäisen sanakirjan avulla, toisaalta käyttäen apuna lauseen muiden sanojen antamia vihjeitä.

Edellisessä esimerkkilauseessa sananmuoto "selässä" tarkoittaa selkän sanan merkitystä "eräs ruumiinos" eikä esimerkiksi merkitystä "ulappa". Koko lauseen kirjaimellinen merkitys selviää lauserakenteesta. Kyseessä on omistusrakenne, jossa ensimmäinen adverbiaali "tytöllä" ilmoittaa omistajaksi nuoren naispuolisen olennon, subjekti "reppu" omistettavan ja jälkimmäinen adverbiaali "selässä" omistettavan esineen paikaksi erään ruumiinosan.

Rakenteen ja merkityksen tunnistusprosessien ei tarvitse olla peräkkäisiä. Päinvastoin on luultavaa, että ihmisellä ne ovat rinnakkaisia. Muoto- ja lauserakenne toimivat vihjeinä, joiden avulla vastaanottaja pyrkii välittömästi ymmärtämään sanoman ensimmäisestä sanasta alkaen. Jo

tunnistettua käytetään hyväksi etsittäessä tulkintaa sanoman muulle osalle.

Ymmärtämisen edellytyksiä

Ymmärtäminen ei toki ole pelkkää sanojen ja lauseiden kirjaimellisten merkitysten selvittämistä. Usein ymmärtäminen edellyttää vielä epäsuorasti ilmaistujen merkitysten tulkintaa. Esimerkiksi "Onko teillä tulta?" ei ole varsinaisesti kysymys vaan pyyntö.

Edelleen ymmärtäminen edellyttää aiemmin sanotun muistamista, selvilläoloa kielenulkoisesta tilanteesta ja arkitiedon hyväksikäyttöä. Täydellinen ymmärrys syntyy vasta, kun nämä prosessit päättelyn avulla yhdistyvät rakenne- ja merkityshahmottukseen. Asiaa valaisee esimerkki:

"Leena sai syntymäpäivälahjaksi leikkijunan. Hän avasi paketin ennen kuin äiti tuli."

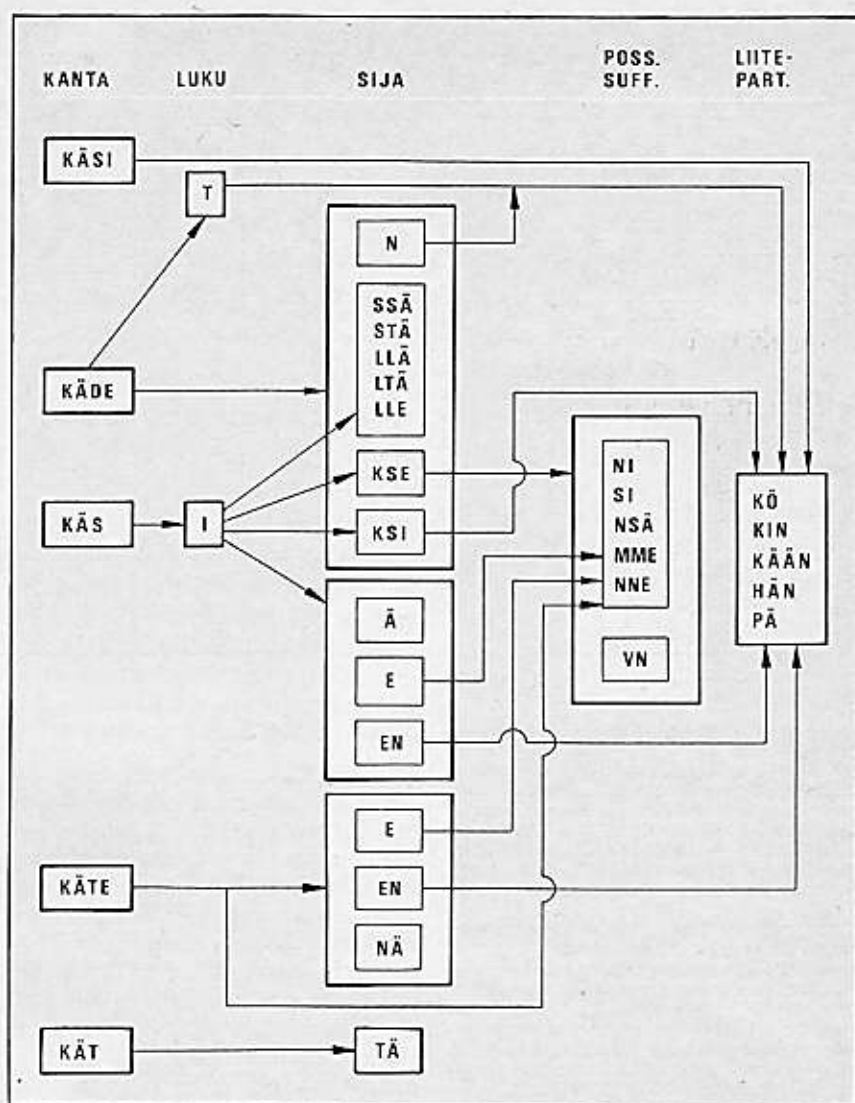
Tämän tekstinpätän ymmärtäminen edellyttää rakenteen ja perusmerkitysten hahmottamisen ohessa mm. sitä, että hän-pronominin tulkitaan tarkoittavan edellisessä lauseessa mainittua Leenaa, ei äitiä. Tämä taas edellyttää kuvaillun tilanteen hahmottamista ja puheena olevien tarkoiteiden muistamista. Toiseksi lahja, leikkijuna ja paketti yhdistetään toisiinsa arkitiedon avulla: leikkijuna on luonnollinen lahja lapselle, ja lahjat annetaan yleensä paketissa (tekstissä ei suoraan sanota, että paketissa oli kyseinen leikkijuna).

Ongelmallinen arkitieto

Edellä esitetyn tarkoituksena on osoittaa, millaisia vaikeuksia on tutkijalla, joka pyrkii muodostamaan malleja kielen ymmärtämistäpahtumasta ja saamaan tietokoneen ymmärtämään luonnollista kieltä.

Erityisen hankalaksi kuvattavaksi on osoittautunut tavallinen arkitieto. Huomattava osa näitä asioita selvittelvästä ns. tekoälytutkimuksesta on kohdistunut arkitiedon ja siihen perustuvan päättelyn kuvaamiseen, mutta silti tässä on otettu vasta ensi askeleet.

Se, että tietokone ymmärtäisi täydellisesti vapaata kirjoitettua kieltä, on vielä kaukaisen tulevaisuuden haave, puheen ymmärtämisestä puhumattakaan. Luultavasti mikään nykyisistä teorioista ei riitä tämän tehtävän ratkaisun pohjaksi. Suuri



osa tarvittavasta perustutkimuksesta on vielä tekemättä. Vaikka kielen rakennemallien ja perusmerkityksen tutkiminen on vain osa koko tehtävästä, se on paitsi teoreettisesti myös käytännön kannalta tärkeää.

Englantia tutkittu eniten

Luonnollisen kielen ymmärtämisen tutkiminen on ennen kaikkea kielitieteellistä perustutkimusta. Kokemus osoittaa, että luonnollisia kieliä käytäviä toimivia tietokonesovelluksia ei saada aikaan ilman kielitieteellisesti perusteltua kielen rakenteen ja merkityksen kuvausta. Oikeat rakennemallit ovat kaikkien ymmärtämismallien välttämätön edellytys.

Toistaiseksi mallintamistutkimus on kohdistunut pääasiassa englannin kieleen, ja se on ollut vilkkainta Yhdysvalloissa. Englannin kielen tyypil-

Suomen substantiivien rakennemalli.

Substantiivin noin 2000 eri taivutusmuotoa voidaan muodostaa liittämällä sanan perusvartaloon halutut päätteet ja päätelyyhdistelmät eri pääteluoista (yksikkö/monikko, sijamuoto, omistusliite ja liitepartikkeli). Nuolet osoittavat pakollisia reittejä; liettyjä päätteitä ei voida yhdistellä keskenään. Liitepartikkeleiden yhdistelmiä, kuten -kohon, -kinko, ei ole otettu huomioon.

lisiä piirteitä ovat mm. sanojen yksinkertainen muotorakenne — vähän päätteitä, vartalonvaihteluja ja erilaisia sananmuotoja — ja melko kiinteä sanajärjestys.

Nämä englannin kielen erikoisominaisuudet ovat vaikuttaneet ehkä liikaakin tutkimukseen. Siitä huolimatta parhaatkään ohjelmat eivät pysty tuottamaan virheettömiä rakennekuvauksia vapaalle kirjoitetulle englannin kielelle.

Yleisen kielitieteen lähtökohtana on kuitenkin näkemys, että eri kielet

KANTA	SANANJOHDON PÄÄTTEET									TAIVUTUSPÄÄTTEET				
	1 V ₁	2 V ₂	3 V ₃	4 PASS	5 S ₁	6 S ₂	7 A ₁	8 A ₂	9 S ₃	10 LUKU	11 SIJA	12 OM	13 LP1	14 LP2
tunne				tta			v		uude		n			
suosi				t			u	imm	uus					
lue	t	utta			mise						ssa	ni	ko	han
nuk	aht	el			u							mme		
ui	skentele						mattom			i	en			
repä	is	t	y				ne			i	den		kin	
suome	nn	utta			ja						lle			
uhra	utu						vais		uuks	i	ssa	an		
nous					u	kas	mais		uute		e	mme	han	
syö					jä	ttäre					lle			
lue	t	utt	eli		jo					i	lta			
järje	st	el			mä		llinen							
kahvi					la									
kahvi							mainen							
kahvi							mais	in						

Suomen nominien päättepaikkajärjestelmä.

Taulukko kuvaa päätteiden mahdollista järjestyksiä vasemmalta oikealle, kun muodostetaan erityisesti substantiiveja ja adjektiiveja. Joka rivillä on yksi johdettu ja taivutettu esimerkkisana. (Huom. Lue aina koko rivi kerrallaan.) Paikat 1-9 ovat johtimia, paikat 10-14 taivutuspäätteitä ja liitteitä varten.

Kantaan voi liittää yksi, kaksi tai joskus kolmekin verbijohdinta (paikat 1-3).

Seuraavana on passiivin tunnus (4), joka yhdistyy parillisin tunnuksen luokasta 7 (esimerkiksi "tunnettava"). Useimmat substantiivien johtimet ovat paikassa 5. Ne, jotka voivat seurata toista substantiivien johdinta, ovat paikassa 6. Useimmat adjektiivien johtimet taas ovat paikassa 7, muutamal paikassa 8. Adjektiivien johtimet ovat melkein järjestään substantiivien johdinten oikealla puolella. Ainoa poikkeus on ominaisuudenimi, joka on paikassa 9 (esimerkiksi "nousukasmaisuus"). Paikoista 7 ja 9 voidaan johdellaessa palata alkuun, jolloin syntyy sellaisiakin johdoksia kuin "oikeudellisuus" ja "järjestelmällistymättämyys".

Johtimia ei yhdessä sanassa ole yleensä enemmän kuin yksi tai kaksi. Taulukko kuvaa päätteiden keskinäistä järjestyksiä. Paikat 10-14 ovat lukua (10), sijoja (11), omistusliitteitä (12) ja liitepartikkeleita (13-14) varten.

perustuvat yhteisiin toimintatapoihin. Suomen tyyppisen, sanarakenteeltaan monimutkaisen ja vapaampaa sanajärjestyksen noudattavan kielen mallintamisessa joudutaan ottamaan mukaan uusia periaatteita. Tätä kautta voidaan päästä lähemmäksi kaikkiin kieliin sopivaa rakenneteoriaa ja aikanaan ehkä parempia tuloksia muidenkin kuin suomen kielen osalta. Siksi suomen kielen perusteellinen tutkiminen on sekä teoreettisesti että käytännössä tärkeää.

Monimuotoinen suomi

Suomen kielen sanarakenteelle on tyyppillistä taivutusmuotojen, johdos-ten, äänenvaihtelujen ja yhdyssanojen runsaus. Tavallisesta substantiivista voidaan tuottaa noin 2 000 eri taivutusmuotoa ja jokaisesta verbistä ainakin 12 000 taivutusmuotoa. Englannin vastaavat luvut ovat 4 ja 4!

Useista sanoista voidaan muodostaa runsaasti johdoksia ja niistä uusia johdoksia, kuten "tiede" — "tietee/llinen" — "tietee/llis/tää" — "tietee/llis/tä/minen" — "tietee/llis/tä/jä" — "tiede/mäinen" — "tiede/mäis/yys" jne. Sanan yleisrakenteessa on 9 paikkaa johtimille. Johdetutkin sanat taipuvat, jolloin päättepaikkoja on substantiivissa kaikkiaan 15.

Taivutus ja johtaminen eivät ole vain valmiiden palasten peräkkäin asettelemista, vaan sanavartalot ja päätteet esiintyvät usein erimuotoisina sen mukaan, mitä aineksia ympäristössä on. Tällaisia vaihteluja on suomessa, kuten muissakin kielissä, erityisen runsaasti juuri yleisimmissä sanoissa. Siksi kielen muotorakenteen yksinkertaistenkin tunnistusmallien on pakko perustua teoreettisesti selkeään perusratkaisuun.

Yhdyssanojen muodostaminen on suomessa erityisen yleistä, suureksi osaksi spontaania ja rakenteellisesti monimuotoista. Esimerkiksi sopivat sanat: "todellisuuden/hahmottamis/kyky", "työn/saanti/mahdollisuus", "pysähtymis/merkin/anto/nappi".

Sanarakenteen mallintaminen onnistunut

Helsingin yliopiston yleisen kielitieteen laitoksen tutkimushankkeessa on yhtenä osatavoitteena ollut suomen kielen sanarakenteen mallintaminen.

Tämä tavoite on saavutettu. Projektin aikana on kehitetty ns. kaksitasomalli, joka soveltuu (luultavasti) minkä tahansa kielen sanarakenteen analysointiin. Samaan tietokoneohjelmaan voidaan syöttää eri kielten taivutus- ym. säännöt, minkä jälkeen ohjelma tunnistaa ja tuottaa halutun kielen sananmuotoja.

Malli on tehokas ja tiettävästi ainoa, joka on kielestä riippumaton. Toistaiseksi sitä on sovellettu suomen kielen lisäksi japanin, romanian, ranskan, englannin ja ruotsin kieleen sekä muinaiskirkkoslaaviin.

Malli ei koostu tavanomaisista peräkkäisistä muunnoksista, vaan se on laadittu rinnakkaiseksi, yksivaiheiseksi prosessiksi. Sanojen äännevaihteluista vastaavat säännöt ovat mallissa yksinkertaisia ns. äärellisinä automaateina. Kukin automaatti käsittelee yhden vaihtelun ja toimii jokseenkin itsenäisesti.

Äärelliset automaattit

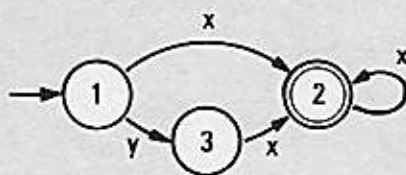
Äärelliset automaattit ovat tunnetuista automaateista yksinkertaisimpia. Niiden mahdollisuus "muistaa" tunnistettavan sarjan aiempia osia on

Äärelliset automaattit

Äärelliset automaattit ovat kuviteltuja koneita, jotka vastaanottavat merkkejä, esimerkiksi kirjaimia. Automaatissa on muutamia tiloja; näitä kuvataan ympyröillä, joiden sisällä on kyseisen tilan nimi tai numero.

Automaatin ajatellaan joka hetki olevan jossakin sen tiloista. Vastaanottaessaan merkkejä automaatti reagoi siirtymällä tilasta, jossa se sillä hetkellä on, uuteen tilaan. Se voi siirtyä myös takaisin samaan tilaan.

Siirtymiä kuvataan nuolilla, jotka osoittavat toisiin tiloihin (tai samaan tilaan). Kunkin nuolen vieressä on merkki (tai vaihtoehdotiset merkit), jonka vastaanottaminen aiheuttaa automaatin siirtymisen nuolen osoittamaan tilaan.



Äärellisiä automaatteja voidaan käyttää merkeistä muodostuvien sarjojen tunnistamiseen, so. automaatti voi erottaa tietynlaiset merkkijonot ja hyväksyä ne samalla kun se hylkää muunlaiset.

Alkaessaan vastaanoton automaatti on alkutilassa, joka on merkitty oheiseen kuvaan tyhjästä tulevalla nuolella. Osa automaatin tiloista on loppu- eli hyväksymistiloja, jotka on merkitty kuvassa kaksinkertaisin ympyröin. Merkkijono tulee hyväksytyksi, jos auto-

maatti on viimeisen merkin jälkeen jossakin lopputilassa.

Automaatti hylkää merkkijonot, joissa ollaan lopuksi muissa kuin lopputiloissa, sekä sellaiset jonot, joita tunnistettaessa vastaanotetulle merkille ei ole ollut sopivaa siirtymänuolta.

Oheinen automaatti hyväksyy esimerkiksi merkkisarjan $x x$ siirtymällä alkutilasta ensin x -nuolta pitkin tilaan 2 ja siitä edelleen samaan tilaan 2. Sen sijaan automaatti hylkää jonon y , koska se siirtyy tilasta 1 y -nuolta pitkin tilaan 3, joka ei ole lopputila. $y y$ puolestaan tulee hyväksytyksi siksi, että tilasta 3 ei lähde mitään y -kirjaimella merkittyä nuolta.

Kaksitasomalli

Nähdäksemme, miten sanojen taipumista voidaan kuvata äärellisten automaattien avulla, tarkastelemme sananmuotoa "laseja". Siinä on kolmen äännevaihtelun jäljet. LASI-vartalon lopun "i" ilmenee "e":nä sanan monikko-muodoissa, monikon "i" ilmenee "j":nä vokaalien välissä ollessaan, ja partitiivin "a" noudattaa vartalon määrittämää vokaalisointua (tässä takavokaalisuutta).

Esitämme "lasi"-sanan sanakirjassa yksinkertaisesti muodossa "lasi", monikon tunnuksen muodossa "I" ja partitiivin päätteiden puolestaan muodossa "A". Monikon tunnuksena käytetään isoa kirjainta erokseksi tavallisista "i"-kirjaimista, koska monikon tunnus toimii sanaa taivutettaessa eri tavoin kuin muut "i":t. Partitiivissa iso kirjain "A" edustaa "a":ta tai "ä":tä vokaalisoinnun mukaan.

Asetamme sanakirjan muodot "lasi", "I" ja "A" ja toisaalta todellisen sananmuodon seuraavanlaisen päällekkäisasetelmaan:

lasiA (sanakirjan mukaiset osaset)

laseja (toteutuva muoto)
Vaihtelut liittyvät tässä asetelmassa kolmeen viimeiseen kirjaimiin, jotka eivät ole samanlaisia sanakirjamuodossa ja toteutuvassa muodossa. Ensinnäkin vartalon lopun "i" toteutuu "e":nä, mikä johtuu siitä että monikon "I" on välittömästi oikealla:

lasiA

laseja

- Vain kolmella lihavoidulla äänneellä on vaikutusta siihen, että "i" toteutuu "e":nä. Sääntö siis on kaavamaisesti:

...iI...

...e...

- Vastaavasti monikon "I" toteutuu "j":nä, jos se joutuu vokaalien väliin, sääntönä:

...I...

...vokjvok...

- Viimeinen sääntö koskee vokaalisointua: partitiivin "A" toteutuu joko "a":na tai "ä":nä edeltävien vokaalien mukaan:

...A...

...takavokaali...a...

Kutakin äänneiden vaihtelua varten voidaan laatia yksinkertaisen äärellisen automaatti, joka valvoo säännön noudattamista. Yksityisten kirjainten sijasta tällainen automaatti ottaa vastaan kirjainpareja. Kussakin parissa on yksi sanakirjatason merkki ja sen alapuolella oleva toteutuvan muodon merkki. Yllä mainitussa esimerkissä olisivat vastaanotettavat parit seuraavia:

l	a	s	i	I	A
l	a	s	e	j	a

Kaksitasomallia voidaan aivan konkreettisesti "ajaa" kumpaankin suuntaan: ohjelma päättelee toteutuneesta sananmuodosta, kuten "laseja", mikä on sallittu sanakirjamuoto ja päätesarja (lasiA), ja toisaalta ohjelma tuottaa sanakirjamuodosta ja päätesarjasta, kuten "lasiA", vastaavan todellisen sananmuodon (laseja). Samat automaattit ohjaavat sekä tunnistusta että tuottamista.

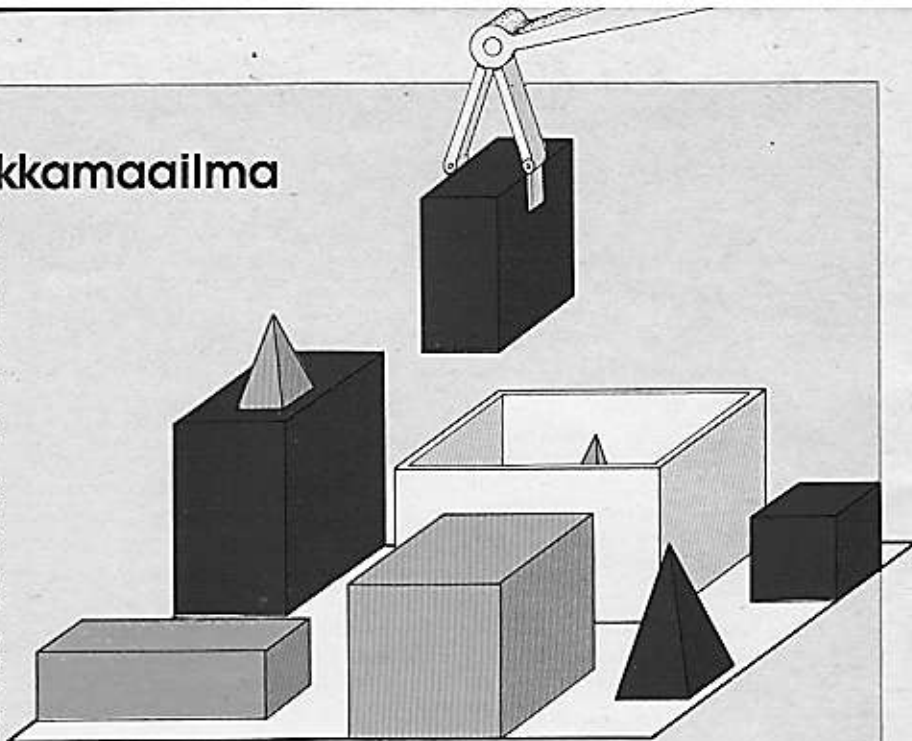
Winogradin palikkamaailma

Ensimmäinen laaja luonnollisen kielen ymmärtämismalli on yhdysvaltalaisen Terry Winogradin kehittämä. Ohjelmisto on tarkoitettu mallintamaan palikoista koostuvaa pienoismaailmaa. Ohjelmisto jäljittelee robottikättä, joka käsittelee erikokoisia ja -värisiä pyramideja ja laatikoita.

Ohjelmistossa on rakenne- ja merkityskomponenttien lisäksi mm. asiayhteyteen perustuva muisti ja palikkamaailmaa koskevia arkitietoja, kuten "pyramidia ei ole mielekästä asettaa kyljelleen", "laatikkoa ei voi asettaa pyramidin päälle", "kun siirtää toisen laatikon alla olevaa laatikkoa, on ensin poistettava päällä oleva laatikko" jne.

Ohessa on keskustelunäyte Winogradin ohjelmiston keskustelu- ja ymmärtämiskyvystä. Numeroidut ilmaukset ovat ohjelmiston käyttäjän syötteksi. Näiden alla ovat järjestelmän vastaukset (kursiivilla). Muutamia kommentteja käydyin keskustelun eri kohtiin:

Noudattaessaan annettua käskyä järjestelmä (tietokone) vastaa OK. Sen pitää olla selvillä, mihin



annetut käskyt viittaavat. Näin ei ole käskyssä 2. "It"-pronominin viittaus edeltävään tekstiin on ratkaistava, sillä pronominilla voi olla useita mahdollisia tulkintoja (3). Vertailuilmausten, kuten "taller", tulkinto on selvitettävä tutkimalla pienoismaailman tilaa (3). Kiellon vaikutusala on otettu huomioon (4).

Eräät käskyt ovat arkitietämyksen vastaisia ja siksi mahdollisia noudattaa (5). Pyramideja ei voi

sijoittaa päällekkäin (6). Järjestelmä päivittää maailmaa koskevia tietojaan (7). Muistin avulla päätellään, mitä kysymys koskee; kaikki aikamuodot ja monet ajan ilmaukset pystytään tulkitsemaan (8,9). Syykysymyksiin ei vastata kaavamaisesti, vaan aktiivisesti pääättelemällä (10, 11 ja 12). Huomaa erityisesti "that"-pronominin erilaiset viittaukset.

- | | | |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| 1. Pick up a big red block.
OK. | 4. How many blocks are not in the box.
<i>Four of them.</i> | 9. When did you pick it up?
<i>While I was stacking up the red cube, a large red block and a large green cube.</i> |
| 2. Grasp the pyramid.
<i>I don't understand which pyramid you mean.</i> | 5. Can the table pick up blocks?
<i>No.</i> | 10. Why?
<i>To get rid of it.</i> |
| 3. Find a block which is taller than the one you are holding and put it into the box.
<i>By "it", I assume you mean the block which is taller than the one I am holding.</i>
OK. | 6. Stack up two pyramids.
<i>I can't.</i> | 11. Why did you do that?
<i>To clear off the red cube.</i> |
| | 7. The blue pyramid is mine.
<i>I understand.</i> | 12. Why did you clear off that cube?
<i>To put it on a large green cube.</i> |
| | 8. Had you touched any pyramid before you put the green one on the little cube?
<i>Yes. The green one.</i> | |

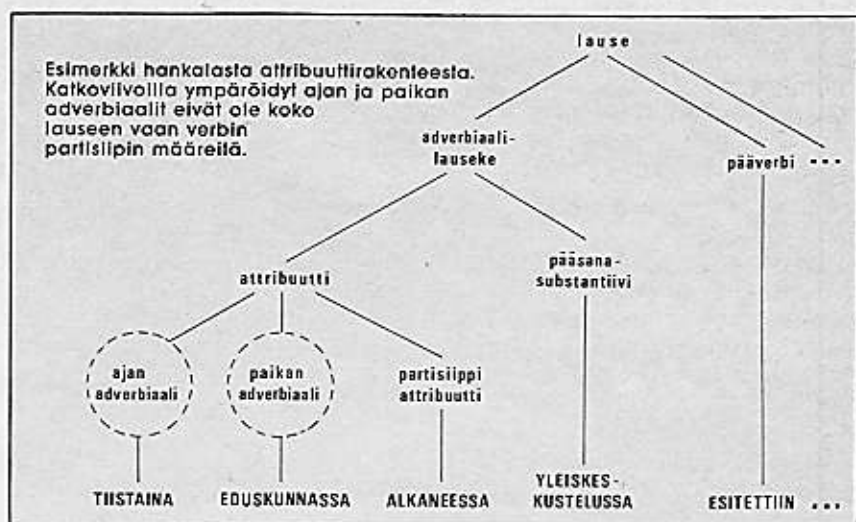
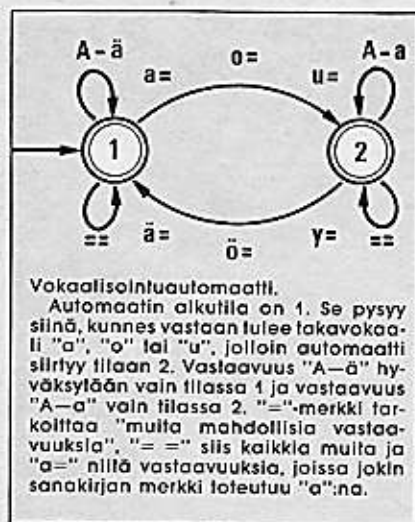
hyvin rajallinen eli ne muistavat ai-noastaan, missä tilassa ovat. Mekanis-mi ei siis muista, miten kyseiseen tilaan tultiin.

Ei ole luultavaa, että ihmisaivoissa olisi osia, jotka noudattaisivat äärellisen automaatin mekanisme sellaise-naan. Asian tekee mielenkiintoiseksi se, että tehtävät, jotka ovat kyllin yk-

sinkertaisia hoidettavaksi äärellisillä automaateilla, voidaan toteuttaa monella muullakin tavalla, esimerkiksi loogisilla elektroniikkapiireillä tai mitä luultavimmin myös aivojen hermosolujen kytkennöillä.

Niinkin monimutkainen ilmiö kuin suomen kielen taivutus ja sanarakenne on kaksitasomallin mukaan

siis pohjimmiltaan hyvin yksinker-tainen, äärellisiin automaateihin rinnastettava mekanismi. Näin tie-tysti täytyykin olla, koska sanojen tai-vuttaminen ja taivutettujen sanojen tunnistaminen sujuvat ilman erityisiä ponnisteluja. Kuitenkin aiemmat sa-nojen taipumista käsittelevät yleiset teoriat edellyttävät olennaisesti ras-



kaampia mekanismeja, jotka olisivat olleet sekä tietokoneille että ihmisäiville ylivoimaisia.

Lauserakenteen tietokonemalli

Suomen kielen koko sanarakenteen toimiva kuvaus ratkaisee melkoisen osan ongelmista, joita mm. englannin kielen lauserakenteen tulkinnassa ilmenee. Suomen sijapäätteet määrittävät usein suoraan, mistä lauseenjäsenestä on kysymys. Päätteet voivat myös näyttää, mikä peräkkäisistä sanoista kuuluvat samaan lauseenjäseneseen. Esimerkissä "tä/ssä uude/ssa ehdotukse/ssa" määritteet mukautuvat pääsanansa sijaan ja lukuun.

Silti suomen lauserakenteen tunnistamisessa on runsaasti ongelmia, esimerkiksi rakenteet, joissa tavallisen substantiivin etumääritteenä on verbistä muodostunut partisiippi ja tällä taas on omina määritteinään adverbiaaleja ja objekteja.

Oheisessa kaaviossa tarkoitetaan siis "tiistaina alkanutta yleiskeskustelua" eikä että "tiistaina esitettiin", kuten adverbiaaleilla yleensä. Näin on esimerkiksi lauseessa "Ehdotus (objekti) esitettiin eilen (adverbiaali) eduskunnassa (adverbiaali)".

Toinen ongelma on nominien -npääte, jolla on varsin monia tehtäviä. Se on milloin subjektin ("Ministeri/n täytyy lähteä"), milloin objektin ("Tapasin ministeri/n"), milloin taas attribuutin ("ministeri/n palkankorotus") merkinä.

Erityisen hankalaa on purkaa rakenteita, joissa on peräkkäin useampia, eri tehtävissä olevia genetiivejä:

"Ministeri tapasi SDP:n (attribuutti) puheenvuoron (objekti) käyttäneen (attribuutti) ryhmän (attribuutti) puheenjohtajan (objekti)..."

Monia sovellutuksia

Luonnollisen kielen ymmärtämisen tietokonemalleilla on monia käytännön sovellusmahdollisuuksia. Jo nyt rakennemalleja kyetään soveltamaan graafisen teollisuuden ladonta- ja toimitusjärjestelmissä sanojen virheettömään tavutukseen, ladottujen tekstien korjauslukuun ja tekstin oikeinkirjoituksen sekä kielenhuolto-näkökohtien tarkistamiseen. Malleja voidaan käyttää myös erilaisissa tiedonhakutehtävissä.

Kielen kääntäminen tietokoneen avulla on teorian puutteesta kärsineen alkunsa jälkeen kehittynyt uudempien rakenneteorioiden varassa huomattavan pitkälle, jopa taloudellisesti kannattavaksi. Parhaat käännösjärjestelmät kääntävät nykyään hyvin tietyn tyyppisiä tekstejä, esimerkiksi teknisiä käsikirjoja.

Kun ihmisen ja koneen käännöksen laatua verrataan tämän tyyppisten tekstien osalta tarkistukseen ja oikaistuihin kuluva aika huomioon ottaen, kone pääsee parempaan tulokseen. Tämä on merkittävä tulos, sillä tekstit käsittelevät sentään niin laajaa aihepiiriä, että mikään varsinainen ymmärtämismalli ei tule kysymykseen.

Tietokone konsultin konsulttina

Tietokonetta voidaan käyttää asiantuntijoiden apuna mm. lääketieteelli-

sessä taudinmäärittämisessä ja öljynporausten tulosten perusteella tehtävässä päätöksenteossa. Tällaisten ns. asiantuntijajärjestelmien menestys ei perustu luovaan "ajatteluun", vaan suoraviivaiseen (pikku)-tarkkaan päättelyyn ja eräänlaiseen "fuskaamiseen" annettujen laajojen säännösten puitteissa.

Nimenomaan asiantuntijasovellusten valtava kysyntä on tehnyt niistä ensimmäisen kaupallisesti kiinnostavan tekoälyntutkimuksen alueen.

Ennen asiantuntijajärjestelmiä oli käytössä tietokantoja, joiden tietoja saattoi kysellä luonnollisella kielellä. Ensimmäinen oli Apollo-projektin kuukivinäytteiden rekisteri 1960-luvun lopulta.

Ihminen tietoyhteiskunnassa

Tietoyhteiskunta tuo mukanaan vakavia ongelmia. Huonoin tapa on antaa tautua "väijäämättömän" kehityksen tai markkinavoimien edessä. Tietokoneiden käyttö on päinvastoin asetettava suuret yksityistä ihmistä palvelevat laatuvaatimukset.

On suuren luokan yhteiskunnallisen kysymys, voivatko kaikki ihmiset käyttää konseisiin tallennettavia tietovarvoja oman äidinkielen avulla vai onko käyttömahdollisuus vain erikoiskoulutuksen saaneilla asiantuntijoilla.

Ongelma koskee myös kansakunnan keskeisiä kulttuuriarvoja. Me emme saa tulla riippuvaisiksi englannin, japanin, venäjän tai jonkin muun kielen ohjelmistoista. Haluamme käyttää kansallisia kieliämme suomea ja ruotsia toimistoissa, tehtaissa, kauppa- ja kotona se-

kä luodessamme tulevaisuuden tietopankkeja tai asiantuntijajärjestelmiä.

Jäämmekö jälkeen?

Yhdysvaltalainen System Development -säätiö on äskettäin sijoittanut 15 miljoonaa dollaria hankkeeseen, jossa useat johtavat amerikkalaiset kielentutkijat ja kielen kysymyksiä pohtivat filosofit tekevät tietokone-asiantuntijain kanssa luonnollisen kielen käsittelyn perustutkimusta. Tavoitteena on teoreettinen läpimurto.

Japanilaisten kohuhanke, viiden sukupolven tietokone, tähtää ratkaisuun, jossa tietokoneen rakenne on suunniteltu nimenomaan tekoälyn ja luonnollisen kielen sovelluksia varten.

Tässä lienee pohdittavaa suomalaisillekin tutkijoille ja tiedepoliitikoille. Tämän tyyppisissä kulttuurin peruskysymyksissä omavaraisuus on eleehto.

(JR)

Aiheesta enemmän

Hakulinen, A. ja Karlsson, F., Nykysuomen lauseoppia. Suomalaisen kirjallisuuden seura, Jyväskylä 1979. 430 s.

Koskenniemi, K., Two-level morphology: A general computational model for word-form recognition and production. Helsingin yliopiston yleisen kielitieteen laitoksen julkaisu 11. Helsinki 1983. 180 s.

Waltz, D. L., Artificial intelligence. Scientific American 247 (1982) 4, s. 101–122.

Winograd, T., Understanding natural language. Edinburgh University Press, Edinburgh 1972. 195 s.

Winograd, T., What does it mean to understand language? Perspectives on cognitive science. Toim. D. A. Norman. Ablex, Norwood, New Jersey 1981. S. 231–263.

Winograd, T., Language as a cognitive process. Vol. 1. Syntax. Addison-Wesley, Reading, Mass. 1983. 610 s.



Pitkä joki

Vuosia sitten tulin ostaneeksi Torontosta 1700-luvun Kanadan kartan, halvan jäljennöksen tosin vain, joka yhä riippuu seinälläni muistona Suuresta Lännestä. Yli kaksi vuosisataa sitten tunnettiin nykyisestä Kanadasta vain itäosa, suunnilleen suurten järvien länsilaidalle asti, ja niinpä kartan vasen laita on jäänyt miltei kokonaan valkoiseksi. Vettä sen nurkkaan kuitenkin on merkitty, vuoria niin ikään, ja vuorten itäpuolella on joki nimeltä Pitkä joki eli Kuollut joki —itse asiassa nykyisten karttojen Missourin yläjuoksu, vaikka se kuvattiinkin Mississippin jatkeeksi. Kartan reunaan on kirjoitettu selitys:

"Pitkän joen eli Kuolleen joen on hiljakkoin löytänyt Paroni de Lahontan, joka on edennyt sitä pitkin karttaan merkittyyn kohtaan asti. Kaikki siitä länteen sijaitseva perustuu piirroksiin, joita gnaksitaarien kansakuntaan kuuluvat villit ovat laatineet hirvennahoilta, ainakin sikäli kuin Herra de Lahontan ei itse ole keksinyt kaikkea tätä, mitä on vaikea arvioida, hän kun on ainoa, joka on tunkeutunut näihin laajoihin erämaihin..."

Kartan laatija Guillaume de l'Isle on epäilemättä ollut perin kriittinen maantieteilijä, ja hän olikin Pariisin kuninkaallisen tiedeakatemian jäsen. Hän selosti ja piirsi uskollisesti kaiken, mikä hänelle oli kerrottu, mutta hän ei halunnut ottaa vastuuta lähteidensä luotettavuudesta vaan antoi päinvastoin lukijansakin ymmärtää, miten vaikeaa, joskaan ei periaatteessa mahdotonta, oli saada kertomuksille vahvistusta puoleen tai toiseen.

Kartan laatijan epäilevyys, etten sanoisi epäluuloisuus, kuvastaa yleisemminkin kaiken tieteen kriittistä henkeä. Tiedehän on itse asiassa

maailman kartoittamista, ikään kuin todellisuuden maantieteen etsimistä, olipa sitten kysymys fyysisestä luonnosta, inhimillisestä yhteiskunnasta tai historiallisesta menneisyydestä. Tällekin kartalle jää aina valkeita alueita. Sillä on myös piirroksia, jotka vastaavat intiaanien hirvennahoilta tekemiä kuvia: kukaan ei tiedä, vastaavatko ne todellisuutta vai ei. Pariisin tiedeakatemiassa asia ei selviä, olivatpa sen filosofit miten syvämielisiä hyvänsä. Ainoa keino on mennä paikalle, koettaa käyttää karttaa ja ottaa selvää, mihin se johtaa. Jokainen suunnistaja kyllä havaitsee, vastaako hänen karttansa todellisuutta.

Kriittinen tutkija tietää, että hänen ainoat apuneuvonsa todellisuutta kartoitettaessa ovat omat silmät ja oma järki — ja hän tietää myös ettei niinkään aina ole luottamista. Hän tietää niin ikään, että ainoa tapa laajentaa kartan tunnetun alueen rajoja on lähteä länteen, pitemmälle kuin kukaan muu, ja ottaa itse selvää, millaiseksi todellisuus siellä voidaan piirtää. Muuta todellisuutta kuin tutkimusmatkailijan kouriintuntuva todellisuus ei tieteellä ole löydettävään, ja vaikka muinoin oli tapana piirtää kartan valkeille laidoille seireenejä, vesinymfejä, kerubeja ja muita taruolentoja, ei se joka meloo tuohikanootissa pitkin Pitkää jokea voi odottaa tapaavansa niitä sen enempää kuin sateenkaaren päässä odottavaa kultamaljaakaan.

Entä kun kaikki on kartoitettu? Päätyykö tutkimus? Tuskinpa vain, sillä yksi vastaus synnyttää kolme uutta kysymystä. Todellisuuden karttaan jää aina, kuten Guillaume de l'Islenkin karttaan, "joki jonka suuta ja lähdeä ei tunneta". Sillä joella tutkija on kotonaan.

Anto Leikola

Tiede 2000 1/1984