

Fred Karlsson



Teoria ja sovellus kielen mallintamisessa

Kognitiotieteen (tai kognitiivisten tieteiden) keskeinen pyrkimys on laatia todennukaisia malleja ihmisen toiminnosta, erityisesti hänen tietoprosesseistaan. Kognitiiviteijälle ei riitä, että hänen rakentamansa malli ulkonaisesti toimii, että sitä voi käyttää työkaluna suorittamaan jotakin tehtävää. Hänen nimenomaisena pyrkimyksenään on mahdollisimman kuvanomaisesti ja täsmällisesti kuvata sitä tapaa, jolla ihminen toimii.

Tästä voidaan päätellä, että kognitiiviteesissä (kuten yleensäkin perustutkimuksessa) keskeistä on *teorian* muodostus. Oikeellinen ja kattava kuvaus jostakin ihmisen tietotoiminnosta edellyttää varmaankin hyvää teoriaa tästä toiminnosta, ja hyvä teoria taas perustuu empiirisissä tieteissä yleensä laajaan havainnointiin. Pelkät havainnot eivät kuitenkaan riitä - sellaisinaan ne helposti hajoavat yksityiskohtien muodottomaksi röykkiöksi.

Hyvälle teorialle tunnusomaista on, että se luo järjestystä ja yleistyksiä. Yksityiskohtat sohteutuvat toisiinsa, ne muodostavat toistuvia osakokonaisuuksia. Keskeinen erotetaan vähemmän keskeisestä ja satunnaisesta, ennustettavat säännönmukaisuudet kuvataan yleistyksillä. Lopulta teorian ytimen muodostavat redusoimattomat keskeiset peruskä-

sitteet, joille on annettu tiukat tulkintasäännöt. Tiedetään täsmällisesti mitä peruskäsitteet merkitsevät ja mitä niistä voidaan johtaa.

Mallilla tarkoitan teorian sovellusta johonkin tiettyyn konkreettiseen ongelmaan, kuvauksen laadintaa tästä ongelmasta. Esimerkiksi Kimmo Koskenniemen tietokonemorfologia, jota esitellään muualla tässä teoksessa, on tarkasti ottaen teoria. Se tarjoaa tutkijan käyttöön joukon yleisiä, hyvin määriteltyjä käsitteitä, joilla voidaan kuvata kielten sanojen taipumista, johtamista ja yhdistämistä sekä näiden ilmiöiden yhteydessä esiintyviä äännevaihteluita. Kysymyksessä on tyypillinen teoria mm. sen takia, että käytetyt käsitteet ovat yleisiä, riippumattomia yksittäisistä kielistä.

Tietyn kielen (kuten swahilin tai ruotsin) sanojen taiputuksen kuvaus kaksitasoteoriaa käyttäen taas on malli, teorian projektio juuri näihin tosiseikkoihin eli tosiseikkojen tulkintaa teorian käsitteiden avulla.

Kognitiotieteen tietointressi eroaa tekoälystä, joka useim-

miten suuntautuu teknisemmin. Tekoälysimulointi tyytyy usein rakentamaan mallin, joka tavalla tai toisella tekee sen mitä kohdekin tekee kiinnittämättä aina paljoo huomiota mallin perustana olevaan teoriaan. Mallin kuvanomaisesta suhdetta kohteeseen ei siis pidetä kovin oleellisena. Tekoälymallit ovat täten usein instrumentaalisempia kuin kognitiiviteelliset mallit.

Tavallinen väärinkäsitys on, että teorian yleensä olisivat jollain kielteisellä tavalla "abstrakteja" tai tehottomia kuvaamaan kohdettaan konkreettisesti. Toki tällaisiakin teorioita on. Haluaisin kuitenkin väittää, että kohdettaan oikeellisesti ja taloudellisesti (mm. mahdollisimman viihin ja yleisin käsittein) kuvaava kognitiivinen teoria on suorastaan välttämätön ennakkoodellytys havainnollisen mallin laadinnalle.

Olemme puhuneet teoriasta ja mallista. Kolmas tärkeä ulottuvuus on *implementaatio*. Kognitiiviteen yhteydessä voi melkein määrittelevänä piirteenä pitää sitä, että tietokonesimuloinnilla on tärkeä osuus, ja sen takia rajoitumme tässä käsittelemään tietokoneimplementointia.

Teorია voidaan implementoida monella eri tavalla, esim. eri ohjelmointikielillä tai samaa ohjelmointikieltä käyttäen eri tavalla eri tietokoneita varten. Huomattakoon siis, että nimenomaan teoria implementoidaan. Mallit (kuten ruotsin taiputuksen kuvaus) toteutetaan käyttäen implementoitua teoriaa. Esim. kaksitasomorfologiasta on monta eri imple-

mentointia, mm. Pascal-, C- ja Prolog-kielillä sekä useilla Lispin murteilla.

Kunnollisen implementaation teko vaatii tietenkin ammattitason atk-tietoja ja -taitoja. Kognitiivisesti relevantin teorian voi toki laatia ilmankin näitä. Todellisuudessa näyttää kuitenkin siltä, että useimmat kognitiotieteilijät laativat teorianakin käyttäen tietojenkäsittelytieteestä lainattuja käsitteitä ja analogioita. Näitä ovat esim. "muisti" (tietyillä tavoilla ymmärretty), "prosessointi", "prossessori", "tiedon siirto", "automaatti" (formaalisten kielten merkityksessä), "top-down" ja "bottom-up" ym. Sen takia puhutaankin usein siitä, että kognitiotiede on tietokone-metaforan läpätunkema - teorian tapa hahmottaa todellisuus nojaa ratkaisevalla tavalla tietokoneen osiin ja sen toimimallisii ominaisuuksiin.

Kognitiotieteen yhtenä leimallisena piirteenä pidetään usein monitieteisyyttä. Tämä onkin tärkeä aspekti. Keskeisenä se toteutuu juuri tietokonesimuloinnin muodossa, jolloin käytännössä yhdistetään esim. filosofian, psykologian tai kielitieteen tietyt lohkot tietojenkäsittelytieteeseen.

Tässä kognitiotiede voi tarjota uusia oivalluksia. Mallin teko implementoidulla teorialla antaa nimittäin kaksi uutta tärkeää mahdollisuutta, johon tavanomainen kynällä ja paperilla tehty teoria ei kykene siitä riippumatta, kuinka syvä ja valaiseva se on. Tietokoneohjelma toimii *ajassa* ja voi siten (parhaassa tapauksessa ja tietyt oivaltavan teorian ohjauksessa) realistisesti mallintaa ajallisia toimintoja - sellaisiahan toiminnat määritelmän mukaan ovat.

Ja toiseksi tietokoneohjelmat tarjoavat mahdollisuuden sekä mallien että niiden pohjana olevan teorian laajaan *testaukseen*. Esim. sanantaivutusmallejamme testamme aina soveltamalla ohjelmistoja satoja tuhansia tai miljoonia sananmuotoja käsitteisiin tekstimassoihin. Tällöin käy aina ilmi, että hyvä-

täkin kuvauksesta on unohtunut joitakin sanaryhmiä tai yksittäisiä sanoja, jotkin sanat saattavat taipua toisin kuin mallin laatija on oletanut (tai toisin kuin normatiiviset kieliooppikirjat ovat väittäneet), todellisesta kielenkäytöstä löytyy runsaasti tapauksia, joista ei ennen oltu edes tietoisia (vaikkapa yleisiä lainasanoja tai lyhenteitä), jne. Mallit korjataan, jonka jälkeen tehdään uusia testejä jälleen uusiin teksteihin, jne.

Metodisena otteena tällainen iteratiivinen työskentelytapa tarjoaa runsaasti uutta. Ennen kaikkea se johtaa huomattavasti tarkempiin eli todennukaisempiin malleihin. Mutta usein iteratiivinen testaus paljastaa aukkoja myös teoriasta. Saataan esim. havaita, että tietynlaisia tapauksia on niin paljon, että jotakin teorian perifeeriseksi luultua osaa pitääkin hioa ja nostaa keskeisemmäksi tai tehdä siitä entistä huomattavasti tarkempi. Pahempi tapaus on sellainen, joka ei ylipäätään sovi kuvattavaksi teorian käsitteillä. Tällöin teoriaa täytyy tietenkin muokata uudelleen niin, että se kattaa myös uudet toiseikat.

Lopputuloksena on, että niin teoriat kuin mallitkin tarkentuvat. Tällainen totuuden asymptootin lähestyminen on empiiriselle tieteelle hyvin tyypillinen. Empiirinen tiede ei tuota lopullisia ikuisia totuuksia, vaan ajan myötä dynaamisesti tarkentuvia kuvia tutkimuskohteesta.

Toistaiseksi olen puhunut teoriasta, malleista ja implementaatiosta yksinomaan tieteellisestä näkökulmasta. Mikä on näiden suhde todellisen elämän ongelmiin, tuotantoon? Miten niitä voi soveltaa käytäntöön?

Yksinkertainen teesini on tämä: hyvät sovellukset edellyttävät hyvää pohjateoriaa, hyvää mallintekoa ja hyvää implementointia. Kaikki nämä kuuluvat itsestään selvästi hyvän tieteen tunnusmerkkeihin. Eli toisin sanoen hyvän tieteellisen tuloksen ja sen käytäntöön soveltamisen välillä ei olekaan sitä juoppaa minkä siellä usein oletetaan olevan.

Tämän olemme useita kertoja nähneet Helsingin yliopiston tietokone-lingvistiikan tutkimyksikössä. Teoreettinen työ on ensisijaisesti ja kaiken perustana, mutta sen pohjalta olemme (usein yhteistyössä erityiskielen edustajien kanssa) kehittäneet tai olemme kehittämissä sovelluksia sellaisilla kielenkäytön alueilla kuin esim. automaattinen tavutus, automaattinen oikoluku, informaation haku ja tietokoneavusteinen kääntäminen. Kohdekielinä ovat yhteydessä tai toisessa olleet suomi, ruotsi, englanti ja venäjä.

1990-luvulle mentäessä tämäntyyppinen kehitys avaa kiintoisia näköaloja. Ajatellaanpa vain, mitä Euroopan integraatioketähdys tulee tuomaan tullessaan. Maiden välinen monikielinen viestintä tulee kasvamaan moninkertaiseksi. On useita ennusteita siitä, että nimenomaan *viestintä* ja sen hallitseminen tulee olemaan 1990-luvun eurooppalaisen tietoteollisuuden todellinen voima. Tässä asiassa eurooppalaisten osaaminen on tunnetusti toista luokkaa kuin Japanin ja Yhdysvaltojen.

Jos näin käy, mm. tietokone-lingvistiikkaan kohdistuva teoreettisen tiedon ja sovellusten kysyntä tulee räjähdysmäisesti kasvamaan. Selvä on, että tämä antaa (tai ainakin sen pitäisi antaa) ajattelemisen aihetta myös yliopistoille. Kuinka merkittäväksi koemme haasteen, mitä teemme sen tyydyttämiseksi ja minkälaiseksi muotoutuu perinteisen lingvistisen tutkimuksen suhde uusiin tehtäviin?