Kirjallisuutta

Sananmuotojen tunnistuksen ja tuoton mallintamista

KIMMO KOSKENNIEMI *Two-level morphology: a general computational model for word-form recognition and production.* University of Helsinki, Department of General Linguistics, Publications 11. Helsinki 1983. 160 s.

Helsingin yliopiston yleisen kielitieteen laitoksessa on ollut vuodesta 1981 käynnissä Suomen Akatemian rahoittama hanke »Suomen kielen automaattinen tunnistaminen ATK:n avulla». Tarkasteltavana oleva Kimmo Koskenniemen väitöskirja on tämän tutkimushankkeen hedelmiä.

Teoksen viidessä luvussa ja kolmessa liitteessä käsitellään seuraavia pääaiheita: morfologisen kaksitasomallin yleisen, kuvattavasta kielestä riippumattoman formalismin esittely; suomen morfologian kaksitasomallinen kuvaus; Pascal-kielisen tietokoneohjelman spesifikaatio. Tarkoituksenani ei ole tarkastella näitä pääkohtia tasaisesti. Puutun vain joihinkin yleisen kielioppiteorian kannalta tärkeisiin kysymyksiin.

Teoreettis-lingvistinen näkökulma ei ehkä tee kaikinpuolista oikeutta kaksitasomallille. Olisi myös mahdollista tarkastella kaksitasomallia filologisena apuvälineenä ja verrata sitä muihin vastaaviin. Vertailu osoittaisi, että kaksitasomalli on edistyneempi kuin ilmeisesti yksikään automaattinen nimikointi- tai lemmausmalli. Tämä näkökulma jää kuitenkin teoksessa täysin taka-alalle. Kaksitasomalli esitellään nimenomaan lingvistisenä mallina.

1. Taustaa

Generatiivista kielioppia on tapana pitää generatiivisena siitä syystä, että se, varsinaisen generoinnin ohella, antaa eksplisiittisen ja testattavissa olevan rakenteenkuvauksen kohdekielen kaikista korrekteista merkkijonoista ja vain näistä. Generatiivisen fonologian formalismi mahdollistaa aksiomaattisilta järjestelmiltä vaadittavan eksplisiittiyden, mutta testattavuus jää paljolti periaatteelliseksi. Käytäntönä on perustaa kuvaus varsin pienen aineiston varaan ja olettaa, että tällaisen sirpalemaisen otoksen generointiin valjastettu säännöstö soveltuu kohdekielen laajamittaiseenkin kuvaamiseen. Suppean aineiston kannalta validiin säännöstöön saattaa kuitenkin kätkeytyä arvaamattomia sivuvaikutuksia, joiden paikantaminen ja korjaaminen on työlästä. Ihmismielen »tietotekniikka» ei ole omiaan laaja-aineistoisen tiedon käsittelyyn, ja näin ollen perinnäinen laatikkotiedostointi saattaa merkitä ajallisesti rajallisten inhimillisten työresurssien tuhlausta. Teoreettinen lingvistiikka tarvitsee metodiseksi tuekseen tietokonelingvistiikkaa.

Tietokonelingvistiikka on kulkenut osaksi omia uriaan. Se on itse saanut vaikutteita kulloisistakin teoreettisen kielitieteen valtavirtauksista, mutta sen antamat vaikutteet ovat toistaiseksi olleet var-

sin vähäiset. Kielitieteen kannalta eivät tietokonelingvistiikan monesti tiettyihin spesifisiin ongelmiin keskittyvät proseduurit ole olleet teoreettisesti kiintoisia. 1970-luvun alusta lähtien tietokonelingvistiikka on kuitenkin saanut yhä ilmeisempää teoreettistakin merkitystä, ja juuri tähän teoreettisesti antoisaan perinteeseen liittyy myös kaksitasomalli. Se sisältää lingvistisesti motivoituja kuvausvälineitä, esittää teoreettisesti kiintoisia väitteitä ja on testattavissa laajalla aineistolla.

Kaksitasomalli on saanut tietokonelingvistiset vaikutteensa lähinnä kahdelta suunnalta. Lauri Karttunen tutkimusryhmineen on kehittänyt suomen kielen sanojen morfologista analyysiä varten TEXFIN-mallin (Karttunen ym. 1981); siitä on saatu idea kuvata morfotaksi ja suppletivismi pienleksikkojen verkkona. Martin Kayn ja Ron Kaplanin kehittelemästä ns. morfografeemisesta kieliopista (Kaplan—Kay 1981; Kay 1983: 100—104) on omaksuttu ajatus, että yksityinen fonologinen sääntö on tulkittavissa ja rakennettavissa äärelliseksi automaatiksi, joka vertaa säännön syötettä ja tulostetta keskenään. Morfografeemisessa kieliopissa peräkkäisiä sääntöjä kuvaavasta automaattisarjasta koostetaan yksi makroautomaatti, jonka tehtäväksi loppujen lopuksi tulee leksikaalisista yksiköistä koostuvan fonologisen syvämuodon ja foneemeista, grafeemeista tms. koostuvan pintamuodon vertailu. Suomen ja muiden morfologisesti monimutkaisten kielten kuvaukseen tarvittavista makroautomaateista tulisi kuitenkin arvaamattoman suuria. Koskenniemen kaksitasomalli on laskennallisesti hyvin lähellä Kaplanin ja Kayn mallia, mutta teoreettinen perusta on toinen. Kaplan ja Kay yrittävät operationaalistaa abstraktia fonologiaa; Koskenniemi liikkuu konkreettisen morfologian linjoilla.

Konkreettisuus toteutetaan kaksitasomallissa nerokkaan yksinkertaisella ratkaisulla: leksikaalinen taso ja pintataso ovat yhtä aikaa edusteilla, eikä ole muita tasoja. Näin ollen ei ole generatiivisen fonologian sääntötekniikan synnyttämiä irreaalisia välitasojakaan. Kaksitasomallin varsinainen uutuus on se, että säännöt toimivat toisistaan riippumattomana rinnakkaiskytköksenä. Generatiivisessa fonologiassa taas »systemaattisfoneettisen» asun määrittelevät säännöt ovat kytköksissä sarjaan. Tosin rinnakkain kytkeytyneiden kaksitasosääntöjen riippumattomuus ei ole täysin ongelmatonta. On hyvin kuviteltavissa tilanne, jossa toinen kaksitasosääntö edellyttää toista (Blåberg 1984), ja tämä rinnastuu generatiivisten sääntöjen feeding-relaatioon.

Generatiivinen fonologia on rakenteeltaan yksisuuntainen; ts. sääntöihin kätkeytyy virtuaalinen tuottamismekanismi. Näin ollen generatiivinen fonologia pystyy mallintamaan joltisenkin luontevasti ainoastaan puhujan toimintaa. Puhujasuuntaisuus onkin lähes jokaisen prosessuaalisesti tulkittavissa olevan mallin ominaisuus, olipa kyse abstraktista tai konkreettisesta ennakkokäsityksestä. Konkreettisuutta edustava Linell jopa argumentoi puhujasuuntaisuuden ontologisen ensisijaisuuden puolesta (1979: 39—46). Prosessuaalisesti yksisuuntaisissa malleissa tuotto- ja tunnistusprosessit ovat erillään, jolloin tunnistus pyrkii osoittautumaan laskennallisesti monimutkaisemmaksi toiminnaksi. Kaksitasomallia voidaan »ajaa» kaksisuuntaisesti. Tämä johtuu siitä, että jokainen kaksitasomallin sääntö on tarkoitettu vastaamaan äärellistä automaattia ja implementaatioon kuuluu algoritmi, joka käyttää näitä automaatteja sekä sananmuotojen tunnistukseen että tuottoon. Kaksitasosääntöjen metodisena tehtävänä on leksikaalisen asun ja pinta-asun vertailu. Säännöt ovat siis erityisiä yhtälöitä, jotka tietty vertailupari joko toteuttaa tai jättää toteuttamatta. Koskenniemen mukaan (s. 10) kaksitasomalli on rakenteellisesti epäsuuntainen ja prosessuaalisesti kaksisuuntainen. Yhtä kaikki kaksitasomallikin näyttää olevan lievästi tuottosuuntainen: tyypillisesti kaksitasosäännöt määrittävät pinta-asuja, mutta ne eivät aina paljasta pinta-asujen kaikkia leksi-

kaalisia vastineita (vrt. Karttunen 1983: 169).

2. Kaksitasomalli

Kaksitasomalli koostuu *lingvistisestä formalismista* (luvut 2 ja 3) ja *tietokoneohjelman spesifikaatiosta* (luku 4).

Kahdeksanosioisesta spesifikaatiosta neljä »ylintä» osiota on tarkoitettu Pascal-kielen ominaisuuksista riippumattomaksi ohjelman loogisen rakenteen kuvaukseksi. Tämän spesifikaation pohjalle onkin jo rakennettu useita INTERLISP-, ZETALISP, NIL-LISP- sekä PROLOG-ohjelmia.

Lingvistinen formalismi koostuu *leksikostosta* ja *säännöstöstä*. Leksikkojärjestelmä rakentuu *(ala)leksikoista* ja *kontinuaatioluokista*. Kullakin »hakusanalla» (juuri, affiksi tai suppletiivi) on, paitsi *muoto* ja *merkitys*, myös tiettyyn alaleksikkoon ohjaava *jatko-osoite*, joka »avautuu» leksikkojärjestelmän kontinuaatio-osastossa. Laajin leksikko on juurileksikko. Esim. sanan *hevosena* juuri on *hevo*:

hevo nen/S "Horse S".

"nen/S" on sen pienleksikon osoite, josta suppletiivialternantit *nen/se* ovat löydettävissä:

LEKSIKKO nen/S
nen K " ";
sE S 123 " ".

Kontinuaatioluokka "K" ilmaisee, että *nen*-ainesta seuraa joko liitepartikkeli tai sananloppu; "S123" taas on sijaleksikkojen S1, S2 ja S3 diagrammaattinen yhteisnimi. Leksikon "nen/S" jäsenillä ei ole varsinaista merkitystä, joten merkityskenttä on tyhjä (" ").

Kaksitasosääntöformalismi on tehty generatiivista vakiomuotoa muistuttavaksi. Sääntöjen yleinen rakenne on

CP "op" LC—RC,

jossa CP = aakkosparin muodostama »vastaavuusosa», "op" = operaattori ja LC = vasen konteksti, RC = oikea konteksti. Kun generatiivisessa fonologiassa CP ilmaistaan toisinkirjoitusformalismilla (a → b), kaksitasomallissa CP ilmaisee vastaavuussuhteen $\binom{a}{b}$; ts. leksikaalista *a*:ta vastaa pinta-*b*). Operaattoreita — ja näin ollen myös sääntötyyppejä — on kolme: *kontekstirajoitesäännön* "=>"-operaattori määrittelee "LC--RC":n ainoaksi ympäristöksi, jossa CP on sallittu; *pintaanpakotussäännön* "<="-operaattori ilmaisee, että ympäristössä "LC—RC" ainoa sallittu pintareaalistuma on CP:n pinta-aakkonen; *yhdistelmäsäännön* "<=>"-operaattori ilmaisee välttämättömän ja riittävän ehdon tietylle korrespondenssiparille. Ympäristöehtojen disjunktiivisuus, optionaalisuus ja iteratiivisuus ilmaistaan generatiivisesta fonologiasta osittain tutulla formalismilla. Myös kulmasulkeet "<>" ovat käytössä, Koskenniemen mukaan »in standard generative fashion» (s. 39). Tämä ei tarkasti pidä paikkaansa. Esim. kaksitasosääntö

$$\overset{a}{\langle B \rangle} \Rightarrow c\text{—}\langle B \rangle d \quad \text{(jossa B = b, p)}$$

yhdistää säännöt

$$\overset{a}{b} \Rightarrow c\text{—}bd \quad \text{ja} \quad \overset{a}{p} \Rightarrow c\text{—}pd,$$

mutta generatiivisen vakiotulkinnan mukaan olisi kyse B-joukon vaihtoehtoisesta puuttumisesta tai mukanaolosta (Sommerstein 1977: 140). Yhtä kaikki kaksitasomalli ei tietenkään ole sidoksissa generatiivisen fonologian käytänteisiin.

Kaksitasosäännöt eivät tätä kirjoitettaessa ole vielä operationaalisia, vaan ne on käsityönä saatettava taulukkoautomaateiksi, joita malli on implementoitu käyttämään. Tämä manuaalinen välivaihe on omiaan siirtämään kuvauksen painopistettä automaatteihin, mikä on lingvistiseltä kannalta nurinkurista. Tekeillä on tiettävästi kaksikin käännösohjelmaa, jotka muuttaisivat kaksitasosäännöt automaattien muotoon. Käännösohjelman pikaista valmistumista ja käyttöönottoa sopii toivoa, sillä sen avulla kaksitasomallin sääntöformalismi itse asiassa muuttuisi morfologiseen ana-

lyysiin erikoistuneeksi korkean tason ohjelmointikieleksi.

Kaksitasokieliopin aakkosto S on kolmikko $<S_s, S_m, S_f>$, jossa S_s on *pinta-aakkosto;* S_m on *morfofoneemisto;* S_f on *morfologinen piirteistö.* Pinta-aakkoset ovat myös leksikon käytössä. Pinta-aakkosin kooditetaan leksikossa myös produktiiviset (ja niin ollen yleiset tai yleistyvät) morfofonologiset vaihtelut (ks. jäljempää).

Morfofoneemit ovat leksikon yksinomaisessa käytössä. Koskenniemi (s. 24) lukee S_m-joukkoon myös prahalaiset arkkifoneemit. Tämä on ongelmallista, sillä arkkifoneemithan eivät ole olemukseltaan morfologisia entiteettejä. Vaikuttaakin siltä, että kyseiset arkkifoneemit ovat itse asiassa sellaisia morfofoneemeja, jotka vain sattuvat osumaan pinta-aakkosten neutraalistumiskohtaan ja ovat niin muodoin tulkittavissa myös arkkifoneemeiksi. Missään tapauksessa ei kyseessä ole järjestelmällinen »arkkifoneemimenettely», jonka mukaan esim. *ranta* ja *räntä* -substantiivien vastaavat leksikaaliset asut olisivat *rantA* ja *räntA*. Koskenniemi käyttää arkkifoneemeja vain, milloin se on morfofonologisesti mielekästä. Esim. sellaisissa sanoissa kuin *saari* sitä ei ole (leksikkojuuri on *saar*, ei *sa:r*). Samoin on vokaalisoinnun laita. »Arkkifoneemeja» näkyy käytettävän vain suffiksaalisessa vokaalisoinnussa — siis tapauksissa, joissa on perimmältään kyse valinnasta kahden pinta-allomorfin (esim. *-na* ja *-nä*) välillä. Juuri tällainen symmetrinen vaihtelusuhde, jossa on mahdotonta tehdä eroa perusalternantin ja johdetun alternantin välillä, on morfofoneemin käytön ihannetapaus (Anttila 1980; Nyman 1982: 23—24; Karlsson 1983: 30—33). Joukko S_m koostukoon vain morfofoneemeista.

Morfofoneemein (tai morfografeemein, jos on kyse kirjoitetun kielen kaksitasokuvauksesta) voidaan siis koodittaa sekä fonologis- että morfologisehtoisia morfeemisegmentin sisäisiä, foneemisegmentin laajuisia vaihteluita. Toisaalta Koskenniemi ei suinkaan koodita kaikkia morfologisehtoisiakaan vaihteluja morfofoneemeiksi. Ratkaiseva kriteeri on hänen mukaansa produktiivisuus (s. 25). Alternaatiotyypin *i—e* (*lasi, lase-*) hän johtaa leksikaalisesta *i*-foneemista ilmentääkseen tämän tyypin produktiivisuutta (s. 74), kun taas vaihtelu *i—e—Ø* (*kivi, kive-, kiv-*) kooditetaan *E*-morfofoneemilla (s. 74—75). Näin muodoin S_m-aakkosto olisi varattu fossiilisempien vaihtelujen kuvaamiseen. Tämä on kiintoisa kriteeri, ja ainakin tekijä itse nojautuu siihen varsin johdonmukaisesti. Tosin lienee vaikea välttyä ongelmilta. Suffiksaalisen *i*:n etisen pintavaihtelun *a—o* Koskenniemi merkitsee morfofoneemilla Å, koska se ei hänen mukaansa ole yhtä yleinen kuin *a—Ø*, jonka hän johtaa leksikaalisesta foneemista *a* (s. 73). Voidaan toisaalta väittää, että *a—o*-alternaatiolla on osittainen funktionaalinen pohja: »*-a*:n *o*:ksi muuttuminen pelastaa loppuvokaalin kadolta ja samalla säilyttää sointuluokan» (Karlsson 1983: 339). Tällainen näkökohta saattaisi hyvinkin johtaa päinvastaiseen analyysiin, jos pysytellään produktiivisuuskriteerissä, semminkin kun ainakin verbeissä *a—o*-alternaatio näyttää olevan yleisempi (Karlsson 1983: 339—40). Tässä lienee kuitenkin paljolti kyse toisaalta kuvaustyylistä ja toisaalta kehittelyn kohteena olevista ominaisuuksista.

Morfologinen tarkkeisto S_f on kaksitasomallin ongelmallisimpia piirteitä. S_f-aakkoset ohjailevat morfologisten yksikköjen käsittelyä laukaisemalla morfofonologisia prosesseja ja merkitsemällä affiksien valintaominaisuuksia. Lingvistit lienevät kutakuinkin yksimielisiä kielellisen merkin teoreettisesta luonteesta: kielellinen merkki sisältää *muodon* (signifiant; signans jne.) ja *merkityksen* (signifié; signatum jne.). Merkki on siis ainakin kaksiulotteinen entiteetti. Tämän teorian ilmentäminen kuvauskäytännössä on aina ollut hankalaa. TG-kieliopissa leksikaalinen yksikkö kuvataan pariksi <D, C>, jossa D on fonologinen piirteistö (»signifiant») ja C on merkin kieliopillista (myös semanttista) käyttäytymistä ohjaavien piir-

teiden kokoelma (vrt. Chomsky 1965: 84). C-piirteiden kimppu on tapana TG-formalismia noudattavissa kuvauksissa esittää alakkain, ei-segmentaalisesti. Esim. genetiivimuoto *radan* voidaan tuottaa leksikaalisista formatiiveista /rata/ ja /+n/ säännön avulla, joka hyödyntää sitä kieliopillista tietoa, että /+n/ on astevaihtelun laukaiseva suffiksi (vrt. esim. Karlsson 1974: 96):

$$t \rightarrow d \,/\, V\_\_V + [{}^{n}_{+} \text{ astevaihtelu}].$$

Jos määrittelemme, että

$\$ = [+$ astevaihtelu], pääsemme lyhempään formulointiin:

$$t \rightarrow d \,/\, V\_\_V + [{}^{n}_{\$}],$$

joka olisi kaksitasosäännöksi käännettynä seuraava (vrt. Koskenniemi s. 80):

$$\overset{T}{d} \Rightarrow V - \$+n.$$

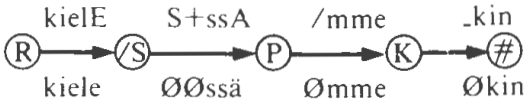
Silmäänpistävä (joskaan ei ilmeisestikään kovin syvällinen) ero on se, että kaksitasomallissa kielelliseen merkkiin kuuluva käyttötieto segmentaalistetaan ja sirotellaan pitkin syntagmaattista akselia. Varsinkin paljon käytettynä tämä on varsin epähygieenistä. Tilanne rinnastuu tekstinkäsittelyohjelmien tulostuskäskyihin, joissa käskytarkkeet tulevat kuvaruutuun varsinaisen tekstin sekaan. Lingvistiseltä kannalta S_f-aakkosten käyttömahdollisuus johtaa metodisiin peruskysymyksiin, sillä nämä tarkkeet ovat käsitteellisesti vallan muuta kuin S_s- ja S_m-aakkoset. Syntagmaattiselle akselille sijoitetut S_f-tarkkeet ovat itse asiassa morfologisen käyttötiedon indeksejä. Tämä on selvästi kaksitasomallin pulmakohtia, kuten Koskenniemi itsekin myöntää (s. 29). S_f-merkkien käyttö tuo kuvaukseen sekä teoreettista että elämyksellistä abstraktiutta, sillä siinä joudutaan turvautumaan verraten suureen joukkoon pinnalla absoluuttisesti neutraalistuvia segmenttejä (suomen kuvauksessa 22) ja lisäksi piirteiden segmentaalistus estää näkemästä, kuinka konkreettisesta kuvausmallista itse asiassa on kyse. Esimerkissä

Leksikaalinen asu:	raTa\$+n
Pinta-asu:	radaØØn

leksikaalis-fonologinen asu sisältää oikeastaan kolmenlaista tietoa, jonka esittäminen segmentaalisena on ongelmallista. T on itse asiassa morfoleksikaalinen, »sanakohtainen» tieto, että kyseinen juuri on astevaihtelun alainen; \$ on foneemisegmentin *n* (»signifiant») käyttöä ohjaava tieto (»signifié» laajassa mielessä), että kyseinen segmentti laukaisee astevaihtelun; + ilmaisee (redundantisti) sen kohdan, mihin yksi morfologisen osajärjestelmän merkki loppuu ja mistä toinen alkaa.

On kuitenkin korostettava, että nämä ongelmat johtuvat viime kädessä kaksitasomallin kiitettävästä eksplisiittisyydestä ja implementaation oikein tai väärin ennakoiduista vaatimuksista. Tämä on joka tapauksessa toista kuin generatiivisen fonologian derivaatioissa ja säännöissä, joissa merkkeihin liittyvä käyttötieto jätetään kuvauksen käyttäjän intuition varaan. Kaksitasomallissakin on optimoitava kahden periaatteen välillä, joista kumpikaan ei ole ihanteellinen. Väitöskirjassaan Koskenniemi turvautuu diakriittien segmentaaliseen käyttöön. Toinen periaatteellinen mahdollisuus on foneemien diakriittinen käyttö. Tämäntapaisesta tasapainoilemisesta kertoneekin se, että kirjassaan (s. 77—78) Koskenniemi koodittaa vokaalien välisen i:n monikollisuuden segmentaalisesti (+i), mutta myöhemmin ilmestyneessä tutkimusta eteenpäin vievässä artikkelissa (Karlsson—Koskenniemi 1985) monikollinen i esitetäänkin »monikollisena morfofoneemina» *I*. Tämäntapainen foneemin diakriittinen käyttö oli leimallista abstraktille fonologialle, mutta se lienee kaksitasomallin nykyimplementaation rajoissa ainoa vaihtoehto diakriittien segmentaaliselle käytölle. Ihanne olisi periaatteessa sellainen implementointi, jossa foneemijonojen prosessointia ohjaava tieto olisi teknisesti itse jonosta eriytettynä. Tällainen muunnos onkin tiettävästi kehitteillä.

Rajamerkit tulkitaan S_1:n osajoukoksi, jonka jäsenet portmanteaunomaisesti ilmaisevat sekä morfologisia liitoskohtia että tiettyjä suffiksiluokkia (s. 26). Ainakin suomen kielen kuvauksessa edellinen tehtävä osoittautuu toissijaiseksi. Morfotaksi ilmenee kaksitasoleksikon osoitejärjestelmästä, joka on eksplikoitavissa tilasiirtymäverkoksi. Esim. sanaa *kielessämmekin* kuvaava verkko on seuraava:[1]



Rajamerkkejä tarvitaan niin muodoin vain, mikäli on kirjoitettava sääntöjä, jotka viittaavat niiden koodittamaan tietoon. Niinpä suomen persoonapäätteet (s. 68) liitetään ilman rajamerkin välitystä verbivartaloon. Suomen kuvauksessaan Koskenniemi käyttää kahdeksaa rajamerkkiä, joista miltei jokainen on tietyn leksikon tai tiettyjen leksikkojen edellyttämä. Esim. "/" on possessiivileksikon P kaikkien jäsenten yhteinen nimittäjä, yhteinen »alkukirjain», joka erottaa omistusliitteet muista suffikseista. Tämä on kätevää varsinkin sääntöjä simuloivien automaattien kannalta, jotka toimiakseen vaativat »ohjausmerkkien» sijoittamista syntagmaattiselle akselille.

Kaksoisristin (#) asemaa Koskenniemi olisi kernaasti saanut selvitellä enemmänkin. Toisaalta kaksoisristiä käytetään sellaisen lopputilan merkkinä, josta voidaan siirtyä vain juurileksikkoon. Tässä käytössä # vertautuu pienleksikkoon. Toisaalta automaatit tuntevat #:n vain sananrajan merkkinä. Tästä päällekkäisyydestä ei sinänsä aiheutune sekaannusta. Järjestelmän sisäiseen logiikkaan kuulunee, että jokaisen juuren »alkukirjain» on #. Tämä implisiittinen näkemys ilmenee, paitsi sääntöformalismista (ks. s. 76, sääntö 5b), myös sanan *postina* satunnaisesta leksikkoesityksestä *#posti* + *nA* (s. 77). Koskenniemi ei kuitenkaan kirjoita juurileksikkonsa hakumuotoja #-alkuisiksi, mikä johtunee mm. siitä, että implementaation kannalta ei juurenalkuista #-merkkiä ole olemassa. Tässä lienee ero mallin logiikan ja kontrollin välillä. Mallin toiminnan kannalta katsoen # näyttää olevan varsinaisesti yhdyssanajunktuurin merkki. Niinpä # esiintyy leksikkoasuissa *s#* (esim. *hevos-*), jonka merkitykseksi jostain syystä ilmoitetaan "YKS NOM" (s. 55), ja $+n#, joka on $+n "YKS GEN" -suffiksin yhdyssana-alkuinen variantti (s. 48). Edelleen # esitetään yksikön nominatiivin yhdyssananalkuiseksi sijantunnukseksi (s. 48; vrt. s. 69). Lingvistiseltä kannalta tällaiset rakenteet eivät ole erityisen elegantteja, mutta automaateille sanoma on yksiselitteinen: Siirry tilaan *1*, initiaalista ja jatka juurileksikkoon.

3. Evaluointia

TG-teorian luonteesta historiallisena ilmiönä on todettu, että se on »an *explanans* in search of an *explanandum*» (Esa Itkonen 1985). Tämä kiteytys sopii periaatteessa kaikkiin tulkintahakuisiin lingvistisiin teorioihin. Jokaisella teoreettisella mallilla on omat rakenneominaisuutensa, jotka tekevät sen virtuaalisesti sopivaksi tai sopimattomaksi tiettyyn mallintamistarkoitukseen. Esim. stratifikaatiokielioppi on varhainen verkkomalli, jonka harrastajat ovat ainakin aikoinaan spekuloineet mallinsa neurolingvistisellä todellisuudella tai relevanssilla (ks. Lockwood 1972: 281—286). Tämä ei sinänsä ole järjetöntä, sillä stratifikaatiokieliopille tällainen tulkinta on paljon luontevampi kuin esim. TG-mallille. Generatiivisen fonologian yksisuuntaiset säännöt ovat puolestaan omiaan mallintamaan palautumattomia prosesseja, siis ennen kaikkea kielihistoriallista evoluutiota (Nyman 1982). Abstrakti fonologia

[1] R = juurileksikko; /S = sijaleksikot; P = possessiivileksikko; K = liitepartikkeli-leksikko.

on metodisesti sisäistä rekonstruointia.

Koskenniemen lähtökohtana on se pitkin 1970-lukua toistettu toteamus, että klassisen generatiivisen fonologian kuvausvoima aiheuttaa ylittämättömiä ongelmia kielen prosessointia (tuottamista ja ymmärtämistä) mallinnettaessa. Tavallaan generatiivisen fonologian klassikot (Chomsky—Halle 1968) kompastuivat tässä omaan näppäryyteensä, sillä pyrkiessään laajentamaan mallinsa tuotekuvaa psykologis-prosessuaalisella tulkinnalla he samalla tulivat osoittaneeksi juuri sen saran, jonka kyntämiseen klassinen generatiivinen fonologia ei ole kovinkaan sopiva. Kaksitasomalli on selvästikin virtuaalinen prosessointimalli, ja Koskenniemi pyrkii tulkitsemaan mallin sisäisiä rakenteita afasiologian ja lapsenkielen tarjoaman ulkoisen evidenssin valossa (s. 128—133). Tämä jakso on ymmärrettävästi spekulatiivinen — liittyyhän ulkoisen evidenssin käyttöön melkoisia filologisia ja epistemologisia ongelmia —, mutta Koskenniemen käyttämät tulkintaperiaatteet ovat kiintoisia ja ilmeisen hyödyllisiä kaksitasomallin sisäisten kuvausvaihtoehtojen arvioinnissa.

Koskenniemi esittää kaksitasomallinsa analogisena *performanssimallina* (s. 127—128), jonka on määrä olla isomorfinen psykologisesti implementoidun morfologisen mallin olennaisten rakennepiirteiden kanssa. Kaksitasomalliin liittyvä tietokoneimplementaatio tietysti vain simuloi reaalisen morfologisen komponentin toimintaa, ts. vain mallin ja mallinnettavan tulosteet ovat identtiset. Voidaan myös sanoa, että kaksitasomallin psykologinen realismi on periaatteessa samaa luokkaa kuin vertailevan metodin historiallinen realismi (vrt. Nyman 1982: 41—46). Vaikka kaksitasomalli hyvin luontuukin prosessointimalliksi, ei psykologista tulkintaa ole suinkaan tarkoitettu mallin keskeiseksi ominaisuudeksi. Ihmismielen prosessointistrategioista tiedetään toistaiseksi valitettavan vähän, ja juuri tällä alueella on pettävän helppoa tehdä metodologiasta ontologiaa. Tämän Koskenniemi tiedostaa kiitettävän hyvin moniin muihin prosessuaalisesti suuntautuneisiin lingvisteihin verrattuna.

Teoksen lingvistisesti syvällisin anti on *kaksitasohypoteesi*, jonka mukaan 1) ei ole muita tasoja kuin leksikaalinen ja pintataso; 2) säännöt toimivat rinnakkaiskytköksenä; 3) fonologiset säännöt ovat simuloitavissa äärellisten tilojen automaateilla. On ennenaikaista arvioida, onko kaksitasohypoteesi kaikin puolin riittävä, mutta tähänastiset tutkimukset eivät ole mallia ainakaan kumonneet — pikemmin päinvastoin. Myöskään mikään edellä esitetyistä kriittisistä huomautuksista ei kumoa itse kaksitasohypoteesia. On tosin ilmiöitä, joiden mallintamiseen kaikki lingvistiset teoriat ovat toistaiseksi olleet riittämättömiä. Esim. ei-proportionaaliset analogiat (Hermann 1931), jotka ovat uudelleen tulleet ajankohtaisiksi skeemoina (Bybee—Slobin 1982), ovat vaikeita esittää, eikä kaksitasomalli tuo tässä suhteessa mitään uutta kuvaan.

Jo ennen kirjan valmistumista kaksitasomalli on ollut elävän kiinnostuksen kohteena. Siitä on konkreettisena todisteena Lauri Karttusen toimittama artikkelikokoelma (1983), jossa mallia sovelletaan englantiin, japaniin, ranskaan ja romaniaan. Muitakin erilliskielisiä sovelluksia on ilmestynyt tai ilmestymässä: muinaiskirkkoslaavi (Lindstedt 1984), ruotsi (Blåberg 1984), uuskreikka (Nyman 1984). Luvassa ovat lisäksi mm. klassisen arabian, saamen ja tšeremissin kaksitasomalliset kuvaukset. Kyseessä on siis sekä kansallisesti että kansainvälisesti huomattava saavutus, mutta pelkin autonomislingvistisin kriteerein arvioituna kaksitasomalli tuskin kykenee erottumaan yksiselitteisesti edukseen muista markkinoilla olevista teorioista. Esim. sellaiset ilmiöt kuin infiksaatio, seemiläisten kielten interdigitaatio, reduplikaatio ja prefiksaatio vaativat vielä leksikkojärjestelmän kehittelyä. Kuitenkin mallin tuotto- ja tunnistusominaisuudet tekevät siitä harvinaisen sovelluskelpoisen — varsin-

kin niin pian kuin sääntökääntäjä saadaan käyttöön —, ja juuri tässä käytännönläheisyydessä on mallin elinvoima.

MARTTI NYMAN

LÄHTEET

ANTTILA, R. 1980: Field theory and morphology. — Lingua Posnaniensis 23 s. 15—19.

BLÅBERG, O. 1984: Svensk böjningsmorfologi: en tvånivåbeskrivning. Yleisen kielitieteen pro gradu -tutkielma. Helsingin yliopisto.

BYBEE, J. L. — SLOBIN, D. I. 1982: Rules and schemas in the development and use of the English past tense. — Language 58 s. 265—289.

CHOMSKY, N. A. 1965: Aspects of the theory of syntax. MIT Pr., Cambridge, Mass.

CHOMSKY, N. — HALLE, M. 1968: The sound pattern of English. Harper & Row, New York.

HERMANN, E. 1931: Lautgesetz und Analogie. Weidmann, Berlin.

ITKONEN, ESA 1985: Arvostelu teoksesta Explanation in Linguistics (toim. N. Hornstein ja D. Lightfoot); Longman, London — New York. — Folia Linguistica 18.

KAPLAN, R. M. — KAY, M. 1981: Phonological rules and finite state transducers. ACL:n vuosikokouksessa (New York City) pidetty esitelmä.

KARLSSON, F. 1974: Centrala problem i finskans böjningsmorfologi, morfofonematik och fonologi. Suomi 117: 2. SKS, Helsinki.

—— 1983: Suomen yleiskielen äänne- ja muotorakenne. WSOY, Helsinki.

KARLSSON, F. — KOSKENNIEMI, K. 1985: A process model of morphology and lexicon. — Folia Linguistica 18.

KARTTUNEN, L. 1983: KIMMO: a general morphological processor. — Texas Linguistic Forum 22 s. 165—186.

KARTTUNEN, L. — ROOT, R. — USZKOREIT, H. 1981: TEXFIN: morphological analysis of Finnish by computer. SASS:n 71. vuosikokouksessa (Albuquerque, New Mexico) pidetty esitelmä.

KARTTUNEN, L. (toim.) 1983: KIMMO: a general morphological analyzer. Texas Linguistic Forum 22 s. 217—228.

KAY, M. 1983: When meta-rules are not meta-rules. — Automatic natural language parsing (toim. K. Sparck Jones ja Y. Wilks; Ellis Horwood, Chichester) s. 94—116.

LINDSTEDT, J. 1984: A two-level description of Old Church Slavonic morphology. — Scando—Slavica.

LINELL, P. 1979: Psychological reality in phonology. Cambridge University Press, Cambridge.

LOCKWOOD, D. G. 1972: Introduction to stratificational linguistics. Harcourt Brace Jovanovich, New York (jne.).

NYMAN, M. 1982: Relational and reconstructive aspects of grammatical systematization; data-oriented studies. University of Helsinki, Department of General Linguistics, Publications 8.

—— 1984: A testable model of Modern Greek morphology. Käsikirjoitus. Helsingin yliopisto, yleisen kielitieteen laitos.

SOMMERSTEIN, A. H. 1977: Modern Phonology. Arnold, London.