ROMTWOL
Suomen romanikielen kaksitasokuvaus

Kimmo Granqvist
Kotimaisten kielten tutkimuskeskus

1. Johdanto

Morfofoginen prosessori on yrityisen tärkeä luonnollisen kielen prosessoinnin työkalu, joka ottaa vastaan syötteenä sanan ja tulostaa siitä kieliopillista tai muuntaa sen haluttuun kieliopilliseen muotoon. Ensimmäisessä tapauksessa on kyse tunnistuksessa, jälkimmäisessä generoinnista. Tavallisesti morfologisen prosessori hajottaa syötteenä annetun sanan morfeemeihin, kuten prefikseihin, vartaloon ja suffikseihin.

Morfologinen prosessori kytkeytyy luonnollisen kielen prosessointityökalun leksikkoon, jolloin vältytään kaikkien taivutusmuotojen sijoittamiselta leksikkoon ja leksikosta tulee huomattavasti taloudellisempi. Kun käsittelee morfologinen prosessori tunnistaa ja generoi taivutusmuotoja, tarvitsee leksikkoon sijoittaa ainoastaan vartalo sekä tieto siitä, millaisia morfeemeja siihen voidaan liittää.

Ensimmäinen kielestä riippumaton morfologisen prosessorin malli oli Kimmo Koskenniemen (1983) kaksitasomalli (Two-Level Morphology, TWOL). Kaksitasokuvauksia on tähän mennessä laadittu ainakin seuraaville kielille: suomi, ruotsi, venäjä, saksa, englanti, ranska, viro, tanska, suahili, baski, turkki, japani, korea ja mari. Osittainen kaksitasokuvaus on julkaistu myös mm. nykykreikalle (Sgarbas & Kokkinakis 1995). Antworth (1990) antaa esimerkkejä mm. memdenistä, tagalogista, hanunoosta ja hepreasta.

Tässä esitelmässä esitellään Suomen romanikielen morfologisen prosessorin implementaatio, joka perustuu kaksitasomalliin. Aluksi esittelen lyhyesti kaksitasomallin, sitten lyhyesti Suomen romanikielen morfologiaa. Luku 4 käsittelee leksikon implementointi, joka on tämän esitelmän pääasia. Luvussa 5 kuvataan lyhyesti sääntökomponentteja ja lopulta luvussa 6 sanakielioppia. Lisäksi esittelen osia sääntö- ja leksikkokomponenttitiedostoista sekä sanakieliopista.

2. Kaksitasomalli ja PC-Kimmo

Kaksitasomalli on tarkasteltavasta kielestä riippumaton malli, joka esittää sanan leksikaalisen muodon, syvämuodon, ja pintamuodon vastaavuutena, joka kuvataan äärellisinä automaatteina (Finite-State Automata) ilmaistujen fonologisten sääntöjen avulla toisessa mallin pääkomponenteista, sääntökomponentissa. Mallin toinen pääkomponentti on leksikkokomponentti, joka luettelee sanastojen syvämuodot sekä lisäksi määrittää morfotaksin.

Kaksitasomalli on kiinnostava sekä tietokonelingvistiikan että deskriptiivisen kielitieteen kannalta. Kaksitasomallista on olemassa useita implementaatioita, joista osa on kaupallisia. Ensimmäinen oli Koskenniemen (1983), jota seurasi pian muita kuten ehkä tunnetuin Karttusen ja hänen opiskelijoidensa (Karttunen 1983; Gajek et al., 1983; Dallrymple et al. 1983). PC-Kimmo on ilmainen implementaatio PC:lle, Macintoshille ja Unixille. Alkuperäinen PC-Kimmo (Antworth 1990) on käsitti klassisen kaksitasokuvauksen, mutta vuoden 1993 toinen versio sisältää laajennuksena sanakieliopin. ROMTWOL on laadittu PC-Kimmo ympäristön versiota 2.1.8 toukokuulta 2000.

Leksikkokomponentti edustaa kielen morfotaktista kuvausta. Kaksitasomallin leksikossa on joukko morfeemien syvämuotoja sekä tieto siitä, millä tavoin niitä voidaan liittää peräkkäin. Morfeemit jaetaan alileksikoihin siten, että samalla tavoin käyttäytyvät morfeemit ryhmitetään yhteen. Lisäksi voidaan määrittää indeksejä (alternations), jotka viittavat alileksikoiden ryhmään tai vain yhteen alileksikkoon. Esimerkiksi ROMTWOLin leksikossa on määrittely `ALTERNATION ROot N AJ V`, joka yhdistää substantiivit, adjektiivit ja verbit Root-indeksin alle. Root-indeksiä hyödynnetään, kun halutaan kertoa, että prefiksejä voi seurata joko substantiivi, adjektiivi tai verbi. PC-Kimmo-ympäristössä glossikenttien avulla voidaan esittää morfeemeista erilaista lisätietoa, esimerkiksi käännöksen ja kieliopillista tietoa.

Kaksitasomallin sääntökomponentti vastaa kielen morfofoneemista kuvausta, joka kertoo, millä tavoin kunkin morfeemin syvämuoto voi muuttua äänneympäristön mukaan. Sääntökomponentti sisältää kielen aakkoston, ts. luettelon syvä- ja pintamuodoissa hyväksyttävistä merkeistä, kolmen erikoissymbolin (NULL, ANY ja BOUNDARY) määrittelyt, aakkosten alaryhmien (subsets) määrittelyt sekä luonnollisesti kaksitasosäännöt. Kaksitasomallin sääntöformalismi muistuttaa generatiivista fonologiaa (Chomsky & Halle 1968), mutta formalismien välillä on merkittäviä periaatteellisia eroja. Siinä missä generatiivisen kieliopin säännöt ovat kontekstista riippuvaisia uudelleenkirjoitussääntöjä formalismien välillä., kaksitasomallin säännöt ovat deklaratiivisia ja ilmaisevat vastaavuuksia syvä- ja pintamuotojen välillä. Tästä seuraa myös toinen ero: toisin kuin generatiivisen fonologian säännöt, joiden soveltamisjärjestys on olennainen, kaksitasosäännöt toimivat rinnakkaisesti. Siitä, että kaksitasosäännöt kuvaavat nimenomaan syvä- ja pintamuotojen vastaavuuksia, johtuu kolmaskin tärkeä ero: kaksitasosäännöt ovat kaksisuuntaisia eli niitä voidaan myös soveltaa pintamuotojen generoimiseen syvämuodoista. Vastaavuuksia on kaksitasomallissa kahta tyyppiä, oletuksia ja erityisiä vastaavuuksia (special correspondences), jotka yhdessä muodostavat sallittujen äänneparien joukon. Säännöt jaetaan perinteisesti neljään eri tyyppiin, joita ilmaistaan formaalissa logiikassa käytettyjen kaltaisilla operaattoreilla. Tietokonetta varten säännöt implementoidaan äärellisiksi automaatteiksi (Finite-State Automata); sääntöjen implementointi tietokoneen lukemaan muotoon käsin on työlästä ja siksi ROMTWOLia laadittaessa apuna käytettiin kgen-sääntökääntäjää (Miles 1991).

Alkuperäinen kaksitasomalli on toimiva jaotelleessaan sanoja tagatuiksi morfeemeiksi, muttei kykene suoraan määrittämään sanan tai sen taivutuskategoiden sanaluokkaa. Tämän rajoituksen vuoksi se ei myöskään sellaisenaan ole riittävä voidakseen toimia syntaktisen parserin morfologisena edustakoneena, jollaisena sillä olisi ehkä eniten käyttöä. Vuonna 1993 kehitetyssä PC-Kimmo-ympäristön versiossa 2.0 puute korjattiin lisäämällä

kolmas analyyttinen komponentti, sanakielioppi, joka on unifikaatiopohjainen PATR-II-formalismiin (Schieber 1986) nojautuva parseri. Parseri tuottaa piirrerakenteita ja puukuvaajia, joissa lehtinä ovat morfeemit. Yksi tärkeimpiä sanakieliopin etuja on se, että se kykenee päättelemään monimutkaistenkin johdosten sanaluokat. Sanakielikomponentti käyttää käyttäjän kirjoittamaa kielioppitiedostoa, joka sisältää piirrerakenteita ja kontekstivapaita sääntöjä. Leksikkokomponenttiin liitetään lisäksi sanakielioppia varten tarvittaessa piirteiden (features) avulla kieliopillisia tietoja morfeemeista, erityisesti affikseista.

3. Suomen romanikielen morfologiaa

Suomen romanikielen aakkosto koostuu 26 kirjainmerkistä (Mustalaiskielen ortografiatoimikunta 1971). Vokaalimerkkejä on kahdeksan, e, i, o, u, y, ä, ö, ja konsonantteja 18: b, d, f, g, h, ђ, j, k, l, m, n, p, r, s, š, t, v, ž. Grafeemia <ђ> käytetään soinnittoman velaarisen spirantin [x] merkkinä. Š ja ž esiintyvät pelkästään affrikaatoissa, joita merkitään <tš> ja <dž>. Aspiroituja klusiileja vastaavat grafeemikombinaatiot <ph>, <th> ja <kh>.

Suomen romanikielessä ei käytetä mitään aksenttimerkkejä, mutta vokaalien ja konsonanttien kestoa, joka on periaatteessa distinktiivinen, merkitään samojen periaatteiden mukaisesti kuin suomen ortografiassakin, esim. <bar> ’markka’: <baar> ’kivi’; <ђinavaa> ’ulostaa’: <ђinnavaa> ’kupata’. Vokaalien ja konsonanttien kestoon vaikuttavat useat morfofonologiset prosessit, joita laukaisevat lähinnä morfeemien konkatenaatiosta aiheutuvat rikkomukset tavurakenteen hyvämuotoisuuteen kohdistuvia rajoituksia vastaan.

Kuten muissakin romanikielen murteissa, Suomen romanikielessä on monimutkainen taivutusjärjestelmä.

Nominien (substantiivien, adjektiivien, pronominien ja numeraalien) kielioppiset kategoriat käsittävät luvun (yksikön ja monikon), suvun (maskuliinin ja feminiinin) sekä sijamuodot. Kieliopillisen suku on ehkä suomen vaikutuksesta suurimmaksi osaksi – joskaan ei missään tapauksessa täysin – hävinnyt, mutta substantiivien taivutukseen edelleen noudattavat vanhoja maskuliini- ja feminiiniparadigmoja. Sijajärjetelmä on muiden romanikielen murteiden tapaan kaksitasoinen ja koostuu toisaalta primaarisista sijamuodoista, toisaalta obliikvisijoista. Primaarisia sijamuotoja on kaksi, nominatiivi ja obliikvi – vokatiivi on hävinnyt kokonaan. Obliikvisijat genetiivi, datiivi, ablatiivi ja instrumentaali muodostetaan suffiksien avulla obliikvista. Lokatiivin käyttö on useimmissa idiolekteissä korkeintaan marginaalista ja rajoittuu muutamiin leksikaalistuneisiin ilmauksiin kuten *drommeste* ’maanteitse’. Elollinen/eloton-kategoria ilmaistaan muiden romanikielen murteiden tapaan morfologisesti akkusatiivissa, joka on joko nominatiivin (eloton) tai obliikvin näköinen (elollinen). Genetiivillä on erityisiä adjektiivinomaisia funktioita, jotka muistuttavat adjektiiviattribuutteja. Koska periaatteesa yhdyssanoja ei ole, genetiivi on tärkeä uusien lekseemien luomisen kannalta, jotka usein muodostetaan genetiivi-adjektiivi + substantiivi –sanaliittoina/lausekkeina.

Substantiivilla on 16 taivutusmuotoa, kun elollinen ja eloton akkusatiivi sekä obliikvi lasketaan mukaan erillisinä taivutusmuotoina. Toisin kuin substantiivit, joilla on yleensä vain yksi kieliopillinen suku, useimmat adjektiivit taipuvat luvun lisäksi myös suvussa, joskin feminiinissä esiintyy vain yksi oma muoto, yksikön nominatiivi. Muut adjektiivien muodot, koko monikko, ovat yhteisiä. Esimerkiksi, adjektiivi ’punainen’ ilmaistaan yksikön nominatiivissa *loolo* (maskuliini), *looli* (feminiini). Poikkeuksen muodostavat kon-

sonanttiin päättyvät alkuperäiset tai lainatut adjektiivit, joilla on defektiivinen paradigma: niillä ei usein ole erillisiä feminiinimuotoa, mutta ne taipuvat silti ei-attributiivisessa käytössä kuten substantiivitkin. Tunnuksellisten substantiivien (alkuperäinen) suku on useimmiten pääteltävissä affiksien perusteella, muttei aina. Esimerkiksi jokseenkin kaikki *–i*-loppuiset substantiivit ovat feminiinejä, mutta *džii* 'sydän' voi taipua joko maskuliini- tai feminiiniparadigman mukaan. *Paani* 'vesi' taas taipuu nykykielessä kuten feminiini mutta on suvultaan maskuliini (esim. drommesko paani, Thesleff 1901). Sen sijaan Ø-merkinnällä varustettujen (konsonanttiin päättyvien) substantiivien sukua ei voida päätellä minkään säännön pohjalta, vaan mm. *kaɦt* 'puu' on maskuliini, kun taas *tšimb* 'kieli' on femiinisukuinen.

Saman suvunkaan sisällä kaikki substantiivit ja adjektiivit eivät taivu samoin, vaikka obliikvisijojen päätteet ovatkin kaikille yhteiset. Säännöllisesti taipuvilla substantiiveilla on 12 erilaista taivutuskategoriaa, kahdeksan maskuliineille ja neljä feminiineille. Mitä tulee adjektiiveihin, vaikka päätteet muistuttavatkin substantiiveja, aiheutuu eroja maskuliini- ja feminiinimuotojen muodostuksesta; kaikkiaan adjektiivien positiivimuotojen taivutus-kategorioita on neljä, lisäksi komparatiiveille yksi oma. Komparatiiveilla ei ole erillisiä maskuliini- ja feminiinimuotoja. Lisäksi toissijaiset adverbit johdetaan suffiksien avulla adjektiiveista, positiiveista tai komparatiiveista. Adjektiivilla on 35 eri taivutusmuotoa.

Pronominien taivutus muistuttaa substantiiveja, mutta muodoissa on runsaasti epäsäännöllisyyksiä.

Verbiparadigma käsittää pääluokan, tapaluokat, aikamuodot sekä persoonamuodot. Pääluokista passiivi on joko geneerinen, aktiivin monikon ensimmäisen tai kolmannen persoonan muotojen avulla ilmaistu, tai sitten 'tulla'-verbin avulla analyyttisesti muodostettu. Romanikielen tapaluokat ovat indikatiivi, konditionaali, subjunktiivi ja imperatiivi. Huomattava piirre on infiniitivin puuttuminen samalla tavoin kuin Balkanin kielissä, joskin subjunktiivi pyrkii osassa idiolektejä käyttäytymään infinitiivin tavoin. Primaarisia, synteettisesti muodostettuja aikamuotoja on morfologisesti katsottuna kolme: preesens, joka vastaa muiden romanikielen murteiden futuuria tai pitkää preseensiä; preteriti ja imperfekti. Viimeksi mainittu on kuitenkin menettänyt Suomen romanikielessä kaikki aikamuodon funktionsa ja sitä käytetään konditionaalina. Synteettisten muotojen lisäksi on muista romanikielen murteista poiketen kaksi liittomuotoa, perfekti ja pluskvamperfekti. Vielä Thesleffin (1901) sanakirjasta löytyvät alkuperäisten synteettiset perfekti ja pluskvamperfekti ovat sulautuneet yhteen nykykielen preteritiksi. Partisiippeja on kahta tyyppiä, preteritin vartalosta muodostettuja adjektiivien tavoin taipuvia, sekä taipumattomia *–men*-päätteisiä (< Kr. *–μένος*), jotka muodostetaan preseensin vartalosta.

Suomen romanikielen verbit jaetaan periaatteessa kolmeen luokkaan: 1) alkuperäiset –aa-loppuiset verbit, kuten *tšeer-aa* 'tehdä', *dikk-aa* 'nähdä', *džaan-aa* 'tietää', 2) alkuperäiset tai lainatut *-(j)av-aa*-loppuiset verbit, jotka ovat usein denominaalisia (harvoin deverbaalisia) transitiivi- tai sitten kausatiiviverbejä, esim. *sikj-av-aa* 'opettaa', *ɦleet-av-aa* 'tasoittaa ' < *ɦleet* 'tasainen', *piv-av-aa* 'juottaa' < *piiv-aa* 'juoda' jne. 3) *–(j)uv-aa*-loppuiset intransitiiviverbit (usein deponentit, inkohatiivit, joskus neutraalit verbit), esim. *sik-juv-aa* 'oppia', *bar-juv-aa* 'kasvaa' < *baro* 'suuri'. Ensimmäisen luokat verbit jakautuvat edelleen neljään alaryhmään, jotka eroavat toisistaan teemavokaalin ja preteri-

tin vartalon muodostuksen osalta, joten verbien taivutuskategorioiden määräksi saadaan kuusi. Verbillä on kaikkiaan 26 finiittimuotoa, mutta kun tavallisimmat rinnakkaismuodot otetaan huomioon, erillisten taivutusmuotojen määräksi tulee *-(j)av-aa* ja *–(j)uv-aa*-loppuisilla verbeillä 48 ja ensimmäisen luokan verbeillä 32, lisäksi yksi taipumaton partisiippi sekä taipuva partisiippi, jolla on adjektiivien tavoin 18 muotoa.

Lisäksi Suomen romanikielessä on rikas derivaatiojärjestelmä. Esimerkiksi Valtonen (1972) luettelee 16 suffiksia, joilla substantiiveista johdetaan adjektiiveja. Substantiiveja voidaan puolestaan johtaa eri tavoin sekä toisista substantiiveista, adjektiiveista että verbeistä, harvoin myös numeraaleista. Yleisimpiä ovat nomen agentis –muodostukset alkuperäistä pitkää genetiivin päätettä käyttäen (esim. *kier-o*) ja abstraktisubstantiivit, joita on kahta päätyyppiä -ba ja -ben –päätteisinä. Verbejä on mahdollista muodostaa sekä nomineista että toisista verbeistä joko transitiivijohtimella *-(j)a(C)v-* tai vastaavalla intransitiivijohtimella *–(j)u(C)*v-. Derivaatioon liittyy myös suuri valikoima sideaineksia, joista tavallisimmat ovat *–i-, -j-* ja *–ji-*.

4. Suomen romanikielen leksikon implementaatio

ROMTWOLin lopullinen leksikko koostuu kymmenestä tiedostosta, mutta tätä kirjoitettaessa ROMTWOLin leksikon uudesta versiosta on valmiina vasta kahden ensimmäisen tiedoston muodostama runko sekä yksi testitiedosto. Ensimmäinen tiedosto sisältää luettelon alternaatioista, piirteistä sekä muista leksikkotiedostoista, jotka ladataan romani.lex-tiedoston jälkeen. Alternaatioita on kaikkiaan 47, joista kaikki muut paitsi neljä ensimmäistä on viittavat toisen valmiin tiedoston, affiksileksikon alileksikoihin. Affiksileksikossa on kuvattuna romanikielen kieliopin pääosa, kaikki kieliopilliset morfeemit (yhteensä 283 leksikaalista entryä).

Jokainen leksikaalinen entry sisältää kuusi kenttää: syvämuodon (**\lf**); sen alileksikon nimen johon syvämuoto kuuluu (**\lx**); tiedon siitä millaisia morfeemeja siihin voi liittyä (**\alt**); piirteet (**\fea**) ja glossikentän (**\gl**), joka on optionaalinen. Piirteitä on toistaiseksi käytetty ainoastaan affiksien yhteydessä kuvaamassa niiden kieliopillisia ominaisuuksia.

Testileksikossa (test.lex) on 30 entryä, jotka edustavat Suomen romanikielen tärkeimpiä säännöllisiä taivutustyyppejä ja morfofonologisia vaihteluita: 14 substan-tiivivartaloa, 4 adjektiivivartaloa sekä 12 verbivartaloa, joista osa on epäsäännöllisiä preteritivartaloita. Substantiivit kuuluvat kaikki samaan alileksikkoon N, adjektiivit alileksikkoon AJ ja verbit alileksikkoon V.

Leksikkokomponentin suunnittelufilosofiana on sisällyttää leksikkoon vain vartaloita, mahdollisimman vähän johdoksia. Tällä saavutetaan useita etuja, mm.: 1) koska parseri kykenee tunnistamaan minkä tahansa säännöllisen johdoksen, vaikkei olekaan lueteltu leksikossa, pelkän vartalon lisäämisellä leksikkoon saadaan automaattisesti tunnistettua kaikki siihen monimutkaisemmat saneet, 2) leksikon koko pienenee ja kuluttaa vähemmän levytilaa, 3) morfologinen rakenne saadaan mallinnettua paremmin. Haittapuolena on ennen kaikkea väärien tulosten lisääntyminen, kun epäproduktiivisia johtimia liitetään vartaloihin. Esimerkiksi denominaalisia verbejä on todellisuudessa melko vähän, ja osa niiden imperatiivin yksikön 2. persoonan muoto on pintarakenteessa saman näköinen monien feminiinien yksikön akkusatiivi ja monikon nominatiivi. ROMTWOL olettaa esimerkiksi virheellisesti, että *phenja* ‘siskoa; siskot’ olisi myös verbin **phenjavaa* impe-

ratiivin yksikön 2. persoonan muoto. Yksi ratkaisu tähän on tietenkin se, että "kielioppia" yksinkertaistetaan ja epäproduktiivisia johtimia käyttävät attestoidut sanan luetellaan leksikossa, mutta tämä taas kuormittaa leksikkoa. Toinen ongelma on se, että osa johdoksista on epäsäännöllisiä eikä niitä voida käytännössä generoida kaksitasosääntöjen avulla. Esimerkiksi *dissuvaa* 'näkyä' on *dikkaa*-verbin 'nähdä' johdos, joka on muodostettu alunperin palatalisoituneen muodon *dič'ol* kautta. Pulmia aiheuttaa tietysti myös se, että monien johdosten merkitys on erikoistunut. Esimerkiksi sana *ranniboskiero* < *ranniba* < *rannaa* 'kirjoittaa' merkitsee 'kirjoittajan' lisäksi myös 'kynää' ja 'viilaa'. Pelkkiin vartaloihin tukeutuva leksikko tietenkään voi tällaista tietää. ROMTWOLissa on toistaiseksi otettu konservatiivinen kanta tähän eikä johdoksia ole lähdetty luettelemaan, vaikka glossien oikeellisuus tästä kärsiikiin. Lopulta myös rajanveto produktiivisten ja epäproduktiisten johtimien ja staattisten diakronisten muodostusten välillä on ongelmallinen

Vartaloiden syvämuotoon vaikuttaa olennaisesti toisaalta painon siirtyminen sanan lopusta suomen mukaisesti sanan alkuun sekä toisaalta kvantiteetin distinktiivisyys ja ortografinen merkitseminen Suomen romanikielessä. Suomen romanikielessä sanojen painollinen alkutavu pyrkii olemaan kaksimorainen, joko (C(C))CVC tai sitten (C(C))CVV, kun taas jälkitavujen suosituin rakenne on CV. Raskaita, yli kaksimoraisia tavutaan vältetään eikä niitä esiinnykään kuin n. 3 % sanoista. Tästä seuraa, että varsinkin vartaloiden alkutavun peak, mutta myös coda on keston usein suhteen määrittelemätön (underspecified), ja niiden kesto pintarakenteessa määräytyy kontekstin mukaan. Esimerkiksi sanan **pa.ni* 'vesi' yksikön nominatiivi toteutuu *paa.ni*, mutta yksikön genetiivi *pan.ja.ko*. Sanan *rom* 'romanimies' yksikön genetiivi kuuluu taas *rom.mes.ko*, ei **ro.mes.ko*. ROMTWOLin leksikossa vartalovokaalien ja konsonanttien kestovaihtelut on otettu huomioon käyttämällä lisämerkintöjä 1 ja 2, niin että sanan *paani* vartaloa merkitään *pa1n* ja sanan *rom* vartaloa *rom2*.

Vartaloita leksikkoon liitettäessä on luonnollisesti tiedettävä oikea taivutustyyppi, jotta alternaatio voidaan määrittää. Tämä ei ole aina ole selvää, koska läheskään kaikkien sanojen taivutuksia ei ole dokumentoitu.

Affiksileksikon tämän hetkisessä versiossa määritellään yhteensä 43 kategoriaa, joista enin osa liittyy taivutukseen ja loput derivaatioon. Substantiivien taivutustyyppejä erotetaan sanan alkuperän ja yksikön nominatiivin perusteella 12 (N-INFL1...N-INFL12), joista kahdeksan tyyppiä on maskuliinisukuisia substantiiveja ja neljä feminiinejä varten:

N-INFL1	Alkup. maskuliinit, Ø-merkintä, nom pl *-a*	*rom*	'romani'
N-INFL2	Alkup. maskuliinit, Ø-merkintä, nom pl *-Ø*	*kaȟt*	'puu'
N-INFL3	Alkup. maskuliinit, Ø-merkintä, nom pl *-e*	*tšeer*	'koti'
N-INFL4	Alkup. feminiinit, Ø-merkintä, nom pl *–ja*	*pheen*	'sisko'
N-INFL5	Alkup. feminiinit, Ø-merkintä, nom pl *–a*	*tšimb*	'kieli'
N-INFL6	Alkup. maskuliinit, nom sg *–o*	*raklo*	'poika'
N-INFL7	Alkup. feminiini, nom sg *–i*	*romni*	'romaninainen'
N-INFL8	Lainatut maskuliinit, nom sg *-os*	*stedos*	'paikka'
N-INFL9	Lainatut maskuliinit, nom sg *–is*	*komunis*	'ihminen'
N-INFL10	Lainatut feminiinit, nom sg *–a*	*neura*	'tuohi'
N-INFL11	Johdetut maskuliinit, nom sg *-ba*	*drabiba*	'lukeminen'
N-INFL12	Johdetut maskuliinit, nom sg *-ben*	džaaben	'meno, kulku'

ja adjektiivien neljä, mutta komparatiivi vaatii omat alileksikkonsa, yhden tunnusta *–de-* varten, toisen sijataivutuksille:

A-INFl1	Alkup. tai lainatut adjektiivit, sg *–o,* pl *–e*	*loolo*	‘punainen’
A-INFL2	Alkup. tai lainatut adjektiivit, sg *–o,* pl *–a*	besko	‘pieni’
A-INFL3	Attribuutteina taipumattomat adjektiivit	*kuɦ*	‘kallis’
A-IFFL4	Adjektiivit, joilla on defektiivinen taivutus	*stöt*	‘lyhyt’
A-INFL5	Komparatiivien sijapäätteet		
COMP	Komparatiivin tunnus	*de*	

Kategoriat N-INFL1...N-INFL12 ja A-INFL1...A-INFL5 sisältävät primaaristen sijamuotojen päätteet ja adjektiivien osalta myös adverbin muodostuksessa käytettävät suffiksit. Obliikvisijojen päätteet sen sijaan ovat kaikille yhteisiä ja sijaitsevat omissaan alileksikoissaan OBLIQUE-SG ja OBLIQUE-PL, joista edelliset liittyvät yksikön obliikviin ja jälkimmäisten monikon obliikviin. Romanikielessä genetiivi siitä poikkeuksellinen sijamuoto, että se käyttäytyy monessa suhteessa adjektiivin tavoin. Attribuuttina se kongruoi pääsanansa kanssa suvussa ja luvussa ja saa saman tyyppiset päätteet kuin adjektiivitkin, kuitenkaan genetiivi-adjektiiveja ei voida enää taivuttaa sijamuodoissa kuten adjektiivien substantivoituja muotoja, joten genetiivin päätteet on sijoitettuun omaan alileksikkoonsa. Substantiivi- ja adjektiivijohdoksia varten on kaksi erillistä alileksikkoa N-SUFF4 ja A-SUFF4, joissa on määritelty kaikkiaan 13 yleisintä johdinta.

Verbimuotojen rakenne on seuraava: vartalo + (transitiivi/intransitiivijohdin) + (nfa) + persoonapääte. Transitiivi/intransitiivijohdin esiintyy vain II ja III luokan verbeissä, kuten *rakk-av-aa* ‘puhua’ ja *vandr-uv-aa* ‘vaeltaa’. Nfa (Non-Final Affix, Hancock 1995) puolestaan on preteritin tunnus, jolla useita erilaisia allomorfeja (*-d-*, *-l-*, *-id-* tai *–ad-*, viimeksi mainittu vain partisiipeissa), joiden esiintyminen riippuu osittain verbin vartalon lopusta, muttei kuitenkaan ole läheskään aina pääteltävissä. Esimerkiksi verbin *aan-a*a ‘tuoda’ preteritin syvämuoto on **an-l-jom*, joka toteutuu *anjom*, mutta *d_aan-aa* verbin sama muoto kuuluu *džaan-id-om.* Lisäksi osalla yleisimpiä 1 luokan verbejä preteritin vartalo on epäsäännöllinen, esim. *la-a* ‘ottaa, saada’, mutta preteritissä **li-l-jom* > *liijom*. Juuri tästä syystä I luokan verbit on jaettava leksikossa neljään eri ryhmään, niin että nfa:n vaihtelut ja preteritin vartaloiden epäsäännöllisyydet saadaan otetuksi huomioon. Verbien persoonapäätteitä on neljää tyyppiä: nykyajan muotojen tavallisimmat päätteet ovat alileksikossa V-INFL1, mutta luokkien II ja III verbeillä on lisäksi kontrahoituja muotoja, joissa transitiivi/intransitiivijohdin ja persoonapääte ovat sulautuneet yhteen, esim. *rakk-av-ena* ~ *rakk-ila* ‘puhua-tr-ind.prees.3sg’. Kontrahoidut päätteet ovat omassa alileksikossaan (V-INFL1C). Mennen ajan persoonapäätteitä on kahta tyyppiä, jotka on sijoitettu alileksikoihin V-INFL2 ja V-INFL2B.

Lisäksi romanikieli käyttää runsaasti sideaineksia, kuten *–i-*, *-j-i*, *-j-il-* jne. Sideainekset ovat välttämättömiä komparatiivin, kreikkalaistentyyppistem *–men*-loppuisten partisiippien sekä useimpien abstraktisubtantiivien muodostuksessa adjektiiveista, verbeistä tai toisista substantiiveista, mutta niitä tarvitaan usein myös II ja III luokan verbejä johdettaessa, esim. *lool-* ‘punainen’: *lool-id-e* ‘punaisempi’, *lool-i-ba* ‘punaisuus’, *lool-j-av-aa* ‘punata’, *lool-j-i-m*en ‘punattu’. ROMTWOLin leksikossa sideaineksille on varattu kah-

deksan alileksikkoa, kaksi adjektiiveihin liittyville aineksille, yksi substantiiveille ja neljä verbeille sekä lisäksi yksi kaikille yhteinen.

Kategoriat perustuvat Granqvistin (tulossa) Suomen romanikielen suppeaan deskriptiiviseen kielioppiin.

5. Sääntökomponentti

ROMTWOLin sääntökomponentti (romani.rul) käsittää aakkoston ja 18 kaksitasosääntöä. Aakkosto on jaettu kolmeen osaan: konsonantit (C), vokaalit (Vc) ja muut symbolit (Other). Kaksitasosäännöt kuvataan seuraavassa:

(1-2) Säännöt 1-2 kuvaavat oletus- ja erityset vastaavuudet.

(3) Sääntö 3 on optionaalinen sääntöä, jonka mukaan vokaalista *i* tulee ei-tavuakannattava puolivokaali *j* morfeemirajan ja muun vokaalin jäljessä. Siten esimerkiksi *tšir+ia* → *tšir0ja* → *tširja* 'muurahainen+acc.sg.anim /nom.pl. /acc.pl. inanim'.

(4) Sääntö 4 lisää optionaalisesti epenteettisen *j*:n morfeemirajan jälkeen pinta-muodossakin tavuakannattavan *i*:n ja vokaalin väliin, jolloin yhden CVV-tavurakenteen sijasta tuloksena kaksi luonnollisempaa CV-tavua. Esimerkiksi *but2+ia+k+o* → *butt0ija0k0* → *buttijako* 'työ+obl.sg+gen+masc.sg'.

(5) Sääntö 5 muuttaa morfeemien konkatenaation kautta syntyneen syvärakenteen pitkän *jj*:n vokaalin *i* ja *j*:n yhdistelmäksi, joka on Suomen romanikielen tapa ilmaista pitkää *j*:tä ortografiassa (vrt. Borin & Vuorela 1998). Esimerkiksi *di1+l+j2om* → *di0000ijom* → *diijom* 'antaa+nfa+pret.1sg'.

(6) Sääntö 6 kuvaa *v*:n heikkenemisen vokaaliksi *u* konsonantin edellä ja sanan lopussa. Tästä syystä esimerkiksi vartaloa *džuvl-* 'nainen' vastaa pintarakenteessa *džuul-*. Samoin esimerkiksi *gav* → *gau* 'kylä'.

(7) Sääntö 7 kuvaa puolivokaalin *v* kadon kevyiden tavujen lopussa. Käytännössä syvämuodon *v* elidoituu transitiivi- ja intransitiivijohtimissa, esimerkiksi *rak2+av+d+om* → *rakk0a00d0om* → *rakkadom* 'puhua+tr+nfa+pret.1sg'.

(8) Sääntö 8 on sointiassimilaatiosääntö, joka koskee nfa:n allomorfia *–d-* soinnittomaan konsonanttiin päättyvien vartaloiden jäljessä, joten esimerkiksi *aax2+d+om* → *aax00t0om* → *aaxtom* 'olla-nfa-pret.1sg'.

(9-12) Säännöt 9–12 liittyvät vartalon lopun resonanttien *n* ja *l* katoon syvä- tai pintarakenteen puolivokaalin *j* edellä. Säännöt 9 ja 11 kieltävät resonanttien kadon sellaisen vokaalin jäljessä, jota ei ole määritelty keston suhteen. Tällaisia ympäristöstä riippuen pitkiä tai lyhyitä vokaaleja esiintyy lähinnä vain vartaloissa; mikäli tavussa, johon ne kuuluvat, ei ole koodaa, ovat vokaalit yleensä pitkiä, jos kooda taas on esimerkiksi morfeemien konkatenaation tuloksena, vokaalit ovat useimmiten lyhyitä. Sellaisissa tavuissa, joiden suosituin rakenne on CV, pyritään taas koodaakin välttämään, jolloin resonantit hävitetään puolivokaalin edellä. Painollisissa alkutavuissakin resonantit elidoituvat, jos morfeemien konkatenaation tuloksena on kolmen konsonantin yhtymä: resonantit voivat sijaita joko yhtymän keskellä tai joskus alussa. Seuraavassa on muutamia esimerkkejä: *pa1n+ia+k+o* → *pa0n0ja0k0o* → *panjako* 'vesi+obl.sg+gen+masc.sg'; *komun+jen+ge* →

komu00jen0ge → *komujenge* ’ihminen+obl.pl+dat’; *džuvl+ia+ke* → *džuu00ja0ke* → *džuujake* ’nainen-obl.sg-dat’. Viimeisessä esimerkissä toteutuu myös sääntö 6.

(13) Sääntö 13 kuvaa *s*:n elision yksikön instruktiivin päätteen *–ha* edellä: *rakl+es+ha* → *rakl0e00ha* → *rakleha* ’poika+obl.sg+instr.’

(14-15) Säännöt 14 ja 15 kuvaavat vokaaliyhtymien kontraktioita niissä I luokan verbeissä, joiden teemavokaali on *a*. Sääntö 14 koskee preesensin yksikön ja monikon ensimmäisen persoonan muotoja, sääntö 15 muita persoonamuotoja: *dža+aa* → *dža00a* → *džaa* ’mennä+pres.1sg’, *dža+eha* → *dža00ha* → *džaha* ’mennä+pres.2sg’.

(16-17) Säännöt 16 ja 17 määrittävät erikoissymbolin 1 transformaatiot. Erikoissymboli 1 toteutuu samana kuin sitä edeltävä vokaali, mikäli seuraa vain yksi konsonantti + vokaali tai sanaraja. Sääntö 17 on tarpeen, jotta estetään kolmen perättäisen vokaalin yhtymät pintamuodossa. Esimerkiksi *pa1ni* → *paani* → *paani* ’vesi’, mutta *di1+l+j2om* → *di0000ijom* → *diijom* ’antaa+nfa+pret.1sg’.

(18) Sääntö 18 määrittää erikoissymbolin 2 transformaatiot. Erikoissymboli 2 toteutuu samana kuin sitä edeltävä konsonantti, mutta vain mikäli seuraa vokaali. Esimerkiksi *rom2* → *rom0* → *rom*, mutta *rom2+es+ta* → *romm0es0ta* → *rommesta* ’rom+obl.sg+abl’.

6. Sanakielioppi

Schieberin (1986) PATR-II-formalismia muistuttava sanakielioppi on omana tiedostonaan (romani.grm). Sanakielioppitiedosto alkaa piirteiden lyhenteiden määrittelyllä “Let”-lauseiden avulla. Piirteiden lyhenteitä käytetään toisaalta leksikossa **fea**-kentissä, toisaalta myös itse kielioppitiedostossa. Esimerkiksi nominatiivin lyhennettä nom vastaa piirrerakenne `[head case: NOM]`.

Sanakieliopin toinen osa sisältää kieliopillisten kategorioiden mallineet (category templates), jotka ovat kieliopillisiin kategorioihin kuten substantiiveihin ja adjektiiveihin liittyviä piirrerakenteita. ROMTWOLissa määritellään mm. että substantiivien vartalon sanaluokka on N, luku on oletuksena !SG ja sijamuoto !NOM jne. (huutomerkkinä käytetään oletusarvon merkkinä). Tämä vähentää leksikossa tarvittavan informaation määrää, koska oletusarvoja ei tarvitse toistaa enää affiksileksikossa.

Kolmatena osana ja samalla sanakieliopin sydämenä ovat kieliopilliset säännöt, joihin liittyy piirteitä koskevia rajoituksia (feature constraints). Piirrerajoitus koostuu kahdesta piirrerakenteessta, joiden täytyy yhdistyä toisiinsa. Piirrerajoituksilla on kahtalainen merkitys: toisaalta ne rajaavat kieliopillisen säännön alaa, toisaalta taas siirtävät piirteitä puukuvaajan noodista toiseen. ROMTWOLissa ei ensimmäistä piirrerajoituksen käyttömuotoa ole hyödynnetty, mutta jälkimmäinen sen sijaan on laajalti käytössä, niin että parseri osaa palauttaa monimutkaistenkin johdosten sanaluokan.

Esimerkiksi substantiiveista johdettuja adjektiivja koskee piirrerajoitus `<word_1 head pos> = <Ainfl head pos>`, joka kertoo, että koko sanan sanaluokka on sama adjektiivin taivutuspäätteen sanaluokka. Tämän rajoituksen rajoituksen ansiosta romvalo tulkitaan oikein adjektiiviksi ja romvales myös adverbiksi, vaikka kantasanta rom onkin substantiivi.

7. Johtopäätöksiä

Tässä esitelmässä esiteltiin kaksitasokuvaus ja morfologisen prosessorin implementaatio Suomen romanikielelle. Implementaatio perustuu PC-Kimmo-ympäristön versioon 2.1.8 toukokuulta 2000. Leksikkotiedosto (liite 1), sääntötiedosto (liite 2) ja kielioppitiedosto (liite 3) kuvaavat romanikielen säännöllisesti taipuvat substantiivit, adjektiivit ja verbit, jotka muodostavat taivutusjärjestelmän perustan. Laajempi leksikko ja muut sanaluokat liitetään järjestelmään tulevaisuudessa, samalla parannetaan epäsäännöllisyysten muotojen käsittelyä. Kaiken kaikkiaan leksikossa määritellään 47 kategoriaa, joista suurin osa liittyy taivutuksiin, loput derivaatioon. 18 kaksitasosääntöä,. jotka on implementoitu tietokonetta varten rinnakkain toimivina äärellisinä automaatteina, vastaavat syvämuotojen morfofoneemisista transformaatioista. Tulevaisuudessa sääntökomponttia tullaan laajentamaan niin, että se mahdollistaa puhetun kielen muotojen analyysin, koska korpuksissa edelleen ortografia vaihtelee suuresti ja puhekielen variantteja on paljon mukana.

Liite 1: Romanikielen leksikkokomponentti

Alla esitetään osia romanikielen leksikkokomponentista (tiedostot Romani.lex, Affix. lex ja Test.lex). Leksikkotiedostot ovat liian pitkiä esitettäviksi kokonaisuudessaan.

```
; romani.lex
; belongs to ROWTWOL Version 2.0
; Version 2.0
; (C) Kimmo Granqvist 2002, Research Institute for the Languages of Finland

ALTERNATION Prefix      PREFIX
ALTERNATION Root        N AJ V
ALTERNATION Noun        N
ALTERNATION Adj         AJ

ALTERNATION N-Infl1     N-INFL1
ALTERNATION N-Infl2     N-INFL2
ALTERNATION N-Infl3     N-INFL3
:
:
ALTERNATION Oblique-Sg  OBLIQUE-SG
ALTERNATION Oblique-Pl  OBLIQUE-PL
ALTERNATION Gen-End     GEN-END

ALTERNATION  Bnd-A1 BND-A1
ALTERNATION  Bnd-A2 BND-A2
:
:
ALTERNATION N-Suff4     N-SUFF4
ALTERNATION A-Suff4     A-SUFF4
ALTERNATION V-Suff-5       V-SUFF-5

ALTERNATION Comp        COMP

ALTERNATION V-Infl1     V-INFL1
ALTERNATION  V-Infl1c      V-INFL1C
ALTERNATION V-Infl2     V-INFL2
ALTERNATION V-Infl2b    V-INFL2B

ALTERNATION Nfa-1   NFA-1
ALTERNATION Nfa-2   NFA-2
ALTERNATION Nfa-3   NFA-3
ALTERNATION Nfa-5   NFA-5

ALTERNATION Prt-Gr  PRT-GR

ALTERNATION Clitic  CLITIC
ALTERNATION End     End

FEATURES 1pl 1sg 2pl 2sg 3pl 3sg abl acc aj/aj aj/aj-av aj/av aj/v aj/v-n anim
cnd comp dat fem gen imp nfa inanim ind instr intr masc n/aj n/n n/v n/v-n nom
obl pl pres pret prs prt sg subj tr v/n v/v v/v-aj v/v-n

FIELDCODE lf    U
FIELDCODE lx    L
FIELDCODE alt   A
FIELDCODE fea   F
FIELDCODE gl1   G

INCLUDE affix.lex
INCLUDE test.lex
END
; affix.lex
```

```
; belongs to ROWTWOL Version 2.0
; Version 2.0
; (C) Kimmo Granqvist 2002, Research Institute for the Languages of Finland

; !
; LEXICON INITIAL
; !

\lf 0
\lx INITIAL
\alt Prefix
\gl1

; !
; LEXICON PREFIX
; !

\lf 0
\lx PREFIX
\alt Root
\gl1


\lf bi+
\lx PREFIX
\alt Prefix
\fea
\gl1 +NEG1

; !
; NOMINAL MORPHOLOGY
; !
; !
; NOUNS / PRIMARY CASES
; !

; LEXICON N-INFL1
; ”rom”

\lf +0
\lx N-INFL1
\alt Bnd-N1
\fea n/n nom sg
\gl1 +NomSg

\lf +0
\lx N-INFL1
\alt Clitic
\fea n/n nom sg
\gl1 +NomSg

\lf +0
\lx N-INFL1
\alt Clitic
\fea n/n acc sg inanim
\gl1 +AccSg.Inanim

\lf +es
\lx N-INFL1
\alt Clitic
\fea n/n acc sg anim
\gl1 +AccSg.Anim
```

```
:
;
; !
; ADJECTIVES / PRIMARY CASES
; !

; LEXICON A-INFL1
; ”loolo”

\lf 0
\lx A-INFL1
\alt Bnd-A1
\fea
\gl1

\lf +o
\lx A-INFL1
\alt Clitic
\fea aj/aj masc nom sg
\gl1 +Masc.NomSg

\lf +o
\lx A-INFL1
\alt Clitic
\fea aj/aj masc acc sg inanim
\gl1 +Masc.AccSg.Inanim

\lf +es
\lx A-INFL1
\alt Clitic
\fea aj/aj masc acc sg anim
\gl1 +Masc.AccSg.Anim

\lf +es
\lx A-INFL1
\alt Oblique-Sg
\fea aj/aj masc obl sg
\gl1 +Masc.OblSg

\lf +e
\lx A-INFL1
\alt Clitic
\fea aj/aj nom pl
\gl1 +NomPl
:
:
; !
; LEXICON END
; !

\lf 0
\lx End
\alt #
\fea
\gl1

; End of File
```

```
; test.lex
; belongs to ROWTWOL Version 2.0
; Version 2.0
; (C) Kimmo Granqvist 2002, Research Institute for the Languages of Finland

; NOUNS
; LEXICON N

\lf rom2
\lx N
\alt N-Infl1
\fea
\gl1 Rom

\lf kaxt
\lx N
\alt N-Infl2
\fea
\gl1 tree, wood

\lf tse1r
\lx N
\alt N-Infl3
\fea
\gl1 house, home
:
:
; ADJECTIVES
; LEXICON AJ

\lf lo1l
\lx AJ
\alt A-Infl1
\fea
\gl1 red

\lf besk
\lx AJ
\alt A-Infl2
\fea
\gl1 small
:
:
:
\lf rak2
\lx V
\alt Bnd1-5
\fea
\gl1 speak

\lf vandr
\lx V
\alt Bnd1-5
\fea
\gl1 wander

; End of file
```

Liite 2. Romanikielen sääntökomponentti

Seuraavassa on tiedoston romani.rul sisältö.

```
ALPHABET
     b d f g h j k l m n p r s t v x z a e i o u y ä ö ’ + 1 2
NULL 0
ANY @
BOUNDARY #
SUBSET C   b d f g h j k l m n p r s t v x z
SUBSET Vc   a e i o u y ä ö
SUBSET Other   ’ + 1 2


RULE ” 1. defaults” 1 31
    b d f g h j k l m n p r s t v x z a e i o u y ä ö ’ + 1 2 i @
    b d f g h j k l m n p r s t v x z a e i o u y ä ö 0 0 0 0 j @
 1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1


RULE ” 2. defaults” 1 11
    0 j v v d n l a e s @
    j i u 0 t 0 0 0 0 0 @
 1: 1 1 1 1 1 1 1 1 1 1 1


; Glide formation: a morpheme-initial i become a non-syllabic glide before any
vowel
RULE ” 3. i:j <= +:0 ___ Vc:Vc” 3 5

      i  i  +  Vc   @
      j  @  0  Vc   @
 1:   1  1  2   1   1
 2:   1  3  2   1   1
 3:   1  0  2   0   1


; j-epenthesis: insert an epenthetic j between two vowels to obtain CV syllab-
les
RULE ” 4. 0:j => [ C:C | Vc:Vc ]  +:0 i:i ___ Vc:Vc” 5 6

      0  Vc  +  i  C   @
      j  Vc  0  i  C   @
 1:   0   2  1  2  2   1
 2:   0   2  3  2  2   1
 3:   0   2  1  4  2   1
 4:   5   2  3  2  2   1
 5.   0   2  0  2  0   0

; Convert jj arising through morpheme concatenation into ij
RULE ” 5. j:i <=> ____ 2:j” 3 4

      j  j  2   @
      i  @  j   @
 1:   2  3  1   1
 2.   0  0  1   0
 3:   2  3  0   1
```

```
; v-lenition: v is lenited into a vowel u
RULE " 6. v:u <= ':0 (C*) Vc (1:@) ___ +:0 [C:C | #]" 6 10

      v  v  '  Vc  +  #  C  1  v   @
      u  @  0  Vc  0  #  C  @  v   @
 1:   1  1  2   1  1  1  1  1  1   1
 2:   1  1  2   3  1  1  2  1  2   1
 3:   1  4  2   1  1  1  1  6  4   1
 4:   1  1  2   1  5  1  1  1  1   1
 5:   1  1  2   1  1  0  0  1  0   1
 6:   1  4  2   1  1  1  1  1  4   1

; v-elision: delete morpheme-final v before d or a word-boundary
RULE " 7. v:0 <=> +:0 Vc ___ +:0 (+:0 +:0) [C:C | #]" 11 8

      v  v  +  Vc  #  C  v   @
      0  @  0  Vc  #  C  v   @
 1:   0  1  2   1  1  1  1   1
 2:   0  1  2   3  1  1  1   1
 3:   8  4  2   1  1  1  4   1
 4:   0  1  5   1  1  1  1   1
 5:   0  1  6   3  0  0  0   1
 6:   0  1  7   3  1  1  1   1
 7:   0  1  2   3  0  0  0   1
 8.   0  0  9   0  0  0  0   0
 9.   0  0 10   0  1  1  1   0
10.   0  0 11   0  0  0  0   0
11.   0  0  0   0  1  1  1   0


; voice assimilation: d (in NFA) becomes voiceless after any voiceless conso-
nants in the stem
RULE " 8. d:t <=> [f|h|k|p|s|t|x](2:@) +:0 ___" 4 12

      d  d  x  2  +  t  s  p  k  h  f   @
      t  @  x  @  0  t  s  p  k  h  f   @
 1:   0  1  2  1  1  2  2  2  2  2  2   1
 2:   0  1  2  3  4  2  2  2  2  2  2   1
 3:   0  1  2  1  4  2  2  2  2  2  2   1
 4:   1  0  2  1  1  2  2  2  2  2  2   1

; Do not delete l or n after a vowel that unspecified for length
RULE " 9. l:0 / <= 1 ___ +:0 @:j Vc" 5 6

      1  l  +  @  Vc   @
      1  0  0  j  Vc   @
 1:   2  1  1  1   1   1
 2:   2  3  1  1   1   1
 3:   2  1  4  1   1   1
 4:   2  1  1  5   1   1
 5:   2  1  1  1   0   1



RULE " 10. n:0 /<= 1 ___ +:0 @:j Vc" 5 6

      1  n  +  @  Vc   @
      1  0  0  j  Vc   @
 1:   2  1  1  1   1   1
 2:   2  3  1  1   1   1
 3:   2  1  4  1   1   1
 4:   2  1  1  5   1   1
 5:   2  1  1  1   0   1
```

```
; Elision of l & n: delete l and n of underlying C{l,n}j or V[l,n}j, if the vowel
is specified for length
RULE " 11. l:0 <=>  [ Vc | C | v:u | +:0 ] ___ +:0 [ i:j | j:j | j:i ] (2:j)" 9 11

      l  l  +  i  2  j  j  v  C  Vc  @
      0  @  0  j  j  j  i  u  C  Vc  @
 1:   0  2  2  1  1  2  1  2  2   2  1
 2:   7  3  2  1  1  2  1  2  2   2  1
 3:   7  3  4  1  1  2  1  2  2   2  1
 4:   7  3  2  0  1  0  0  2  2   2  1
 5:   0  2  2  1  0  2  1  2  2   2  1
 6:   7  3  2  1  0  2  1  2  2   2  1
 7.   0  0  8  0  0  0  0  0  0   0  0
 8.   0  0  0  1  0  1  1  0  0   0  0
 9.   0  0  0  0  1  0  0  0  0   0  0


RULE " 12. n:0 <=> [ Vc | C | +:0 ] ___ +:0 [ i:j | j:j ] Vc" 9 8

      n  n  +  i  Vc  j  C  @
      0  @  0  j  Vc  j  C  @
 1:   0  2  2  1   2  2  2  1
 2:   7  3  2  1   2  2  2  1
 3:   7  3  4  1   2  2  2  1
 4:   7  3  2  5   2  6  2  1
 5:   0  2  2  1   0  2  2  1
 6:   7  3  2  1   0  2  2  1
 7.   0  0  8  0   0  0  0  0
 8.   0  0  0  9   0  9  0  0
 9.   0  0  0  0   2  0  0  0


; "Elidate s of the oblique stem before ha in instructive"
RULE " 13. s:0 <=> ____ (+:0) h:h a:a" 7 6

      s  s  +  h  a  @
      0  @  0  h  a  @
 1:   2  5  1  1  1  1
 2.   0  0  3  4  0  0
 3.   0  0  0  4  0  0
 4.   0  0  0  0  1  0
 5:   2  5  6  7  1  1
 6:   2  5  1  7  1  1
 7:   2  5  1  1  0  1


; Vowel contractions in verbs with present in a
RULE " 14. a:0 <=> a:a (1:0) (+:0) ___ a:a" 6 6

      a  a  a  1  +  @
      0  @  a  0  0  @
 1:   0  2  2  1  1  1
 2:   5  4  4  3  6  1
 3:   5  4  4  1  6  1
 4:   5  0  0  3  6  1
 5.   0  2  2  0  0  0
 6:   5  4  4  1  1  1
```

```
RULE ” 15. e:0 <=> a:a (1:0) (+:0) ___ C:C” 6 7

      e  e  a  1  C  +  @
      0  @  a  0  C  0  @
 1:   0  1  2  1  1  1  1
 2:   5  4  2  3  1  6  1
 3:   5  4  2  1  1  6  1
 4:   0  1  2  1  0  1  1
 5.   0  0  0  0  1  0  0
 6:   5  4  2  1  1  1  1

; Vowel length: a vowel unspecified for length becomes long if followed by followed by a CV* syllables or C + word boundary
RULE ” 16. 1:{a,e,i,o,u,y,ä,ö} <= {a,e,i,o,u,y,ä,ö} ___ (+:0) C:C (+:0) [Vc:Vc | #]” 12 22

      1  1  1  1  1  1  1  1  1  a  +  C  #  e  i  o  u  y  ä  ö  Vc  @
      a  @  e  i  o  u  y  ä  ö  a  0  C  #  e  i  o  u  y  ä  ö  Vc  @
 1:   1  1  1  1  1  1  1  1  1  2  1  1  1  6  7  8  9 10 11 11   1  1
 2:   1  3  3  3  3  3  3  3  3  2  1  1  1  6  7  8  9 10 11 11   1  1
 3:   1  1  1  1  1  1  1  1  1  2  4  5  1  6  7  8  9 10 11 11   1  1
 4:   1  1  1  1  1  1  1  1  1  2  1  5  1  6  7  8  9 10 11 11   1  1
 5:   1  1  1  1  1  1  1  1  1  0 12  1  0  0  0  0  0  0  0  0   0  1
 6:   3  3  1  3  3  3  3  3  3  2  1  1  1  6  7  8  9 10 11 11   1  1
 7:   3  3  3  1  3  3  3  3  3  2  1  1  1  6  7  8  9 10 11 11   1  1
 8:   3  3  3  3  1  3  3  3  3  2  1  1  1  6  7  8  9 10 11 11   1  1
 9:   3  3  3  3  3  1  3  3  3  2  1  1  1  6  7  8  9 10 11 11   1  1
10:   3  3  3  3  3  3  1  3  3  2  1  1  1  6  7  8  9 10 11 11   1  1
11:   3  3  3  3  3  3  3  3  3  2  1  1  1  6  7  8  9 10 11 11   1  1
12:   1  1  1  1  1  1  1  1  1  0  1  1  0  0  0  0  0  0  0  0   0  1

; Do not permit three consecutive vowels
RULE ” 17. 1:0 <= Vc ____ (+:0) (C:0) (+:0) C:Vc” 6 7

      1  1  Vc  +  C   C   @
      0  @  Vc  0  0  Vc   @
 1:   1  1   2  1  1   1   1
 2:   1  3   2  1  1   1   1
 3:   1  1   2  4  5   0   1
 4:   1  1   2  6  5   0   1
 5:   1  1   2  6  1   0   1
 6:   1  1   2  1  1   0   1

; Stem-final consonants unspecified for length underdo gemination before a vowel
RULE ” 18. 2:{b,d,f,g,h,j,k,l,m,n,p,r,s,t,v,x,z} <= {b,d,f,g,h,j,k,l,m,n,p,r,s,t,v,x,z}  ___ (+:0) Vc” 20 38

     2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  b  +  Vc  d  f  g  h  j  k  l  m  n  p  r  s  t  v  x  z  @
     b  @  d  f  g  h  j  k  l  m  n  p  r  s  t  v  x  z  b  0  Vc  d  f  g  h  j  k  l  m  n  p  r  s  t  v  x  z  @
 1:  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
 2:  1  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
 3:  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  4   0  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
 4:  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  1   0  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
 5:  3  3  1  3  3  3  3  3  3  3  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
 6:  3  3  3  1  3  3  3  3  3  3  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
 7:  3  3  3  3  1  3  3  3  3  3  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
 8:  3  3  3  3  3  1  3  3  3  3  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
 9:  3  3  3  3  3  3  1  3  3  3  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
10:  3  3  3  3  3  3  3  1  3  3  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
11:  3  3  3  3  3  3  3  3  1  3  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
12:  3  3  3  3  3  3  3  3  3  1  3  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
13:  3  3  3  3  3  3  3  3  3  3  1  3  3  3  3  3  3  2  1   1  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20   1
```

```
14:  3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 2 1  1 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  1
15:  3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 2 1  1 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  1
16:  3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 2 1  1 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  1
17:  3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 2 1  1 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  1
18:  3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 2 1  1 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  1
19:  3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 2 1  1 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  1
20:  3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 2 1  1 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  1
```

END

Liite 3: Romanikielen sanakielioppi

Seuraavassa on osia romanikielen sanakieliopista romani.grm. Kielioppitiedosto on liian laaja esitettäväksi kokonaisuudessaan.

```
; romani.grm
; belongs to ROWTWOL Version 2.0
; Version 2.0
; (C) Kimmo Granqvist 2002, Research Institute for the Languages of Finland

;=======================================================
; Feature template definitions
;=======================================================

;----------------------
;Suffixes: from_pos/pos
;----------------------

LET n/n be [from_pos: N
 base_pos: !N
 head: [pos: N
  number: !SG
 case: !NOM
 animacy: !-]]

LET n/aj be [from_pos: N
 base_pos: !N
 head: [pos: AJ
 aform: !ABS
 number: !SG
 animacy: !-]]
:
:
:

;=======================================================
; Category defaults
;=======================================================

LET N be <cat> = ROOT ;noun
 <head pos> = N
 <head number> = !SG
 <head case> = !NOM
 <head animacy> = !-
 <root_pos> = !N
 <base_pos> = !N
:
:
:
; Oblique cases
RULE
Word_1 -> Word_2 {OBLIQUE-SG / OBLIQUE-PL} (GEN-END) (CLITIC)
 <Word_1 head number> = <Word_2 head number>
 <Word_1 head case> = <OBLIQUE-SG head case>
 <Word_1 head case> = <OBLIQUE-PL head case>
 <Word_1 head aform> = <Word_2 head aform>
 <Word_1 head animacy> = <Word_2 head animacy>
 <Word_1 head pos> = <Word_2 head pos>
 <Word_1 root> = <Word_2 root>
 <Word_1 root_pos> = <Word_2 root_pos>

END
```

Liite 4: Koetuloksia

Seuraavassa esitetään eräitä koetuloksia, jotka on tuotettu PC-Kimmon FILE RECOGNIZE -optiota käyttäen. Aluksi tulokset esitetään ilman kielioppiatiedostoa romani.grm, sitten kielioppitiedostoa käyttäen.

```
rom                         ‘rom2+ Rom+NomSg
                            ’rom2+ Rom+AccSg.Inanim
rommesta                    ’rom2+es+ta Rom+OblSg+Abl
romvales                    ’rom2++val+es Rom+NomSg+AJR8+Masc.AccSg.Anim
                            ’rom2++val+es Rom+NomSg+AJR8+AV1
                            ’rom2++al+es Rom+NomSg+AJR1+Masc.AccSg.Anim
                            ’rom2++al+es Rom+NomSg+AJR1+AV1
tseeresko                   ’tse1r+es++ko house, home+AccSg.Anim+CLIT3
                            ’tse1r+es+k+o house, home+OblSg+Gen+MascSg
                            ’tse1r+es++ko do+SUBJ.PRES.2SG+CLIT3
romjako                     ’rom2++j+av+++ko    Rom+NomSg+j+TR1+IMP.2SG+CLIT3
                            ’romn+ia++ko Romni+AccSg.Anim+CLIT3
                            ’romn+ia+k+o Romni+OblSg+Gen+MascSg
                            ’romn+ia++ko Romni+NomPl+CLIT3
                            ’romn+ia++ko Romni+AccPl.Inanim+CLIT3
dzuujensa                   ’dzuvl+ien+sa Woman+OblPl+Instr
looli                       ’lo1l+i red+Fem.NomSg
loolovitika                 ’lo1l+ovitik+a red+AJR10C+NomPl
                            ’lo1l+ovitik+a red+AJR10C+AccPl.Inanim
                            ’lo1l+ovitik+av+ red+AJR10C+TR1+IMP.2SG
kuxxeske                    ’kux2+es+k+e expensive+OblSg+Gen+Pl
                            ’kux2+es+ke expensive+OblSg+Dat
kuxjula                     ’kux2++j+ula expensive+NomSg+j+INTR.IND.PRES.3SG
tserdom                     ’tse1r+d+om do+NFA1+PRET.1SG
aaxxelas                    ’aax2+elas be+CND.PRES.3SG
rakkimen                    ’rak2+i+men speak+BND1+PRT
rakkibosko                  ’rak2+i+bos+k+o speak+BND1+OblSg+Gen+MascSg
rakkiboskiereske            ’rak2+i+bos+kier+es+k+e
                                        speak+BND1+OblSg+NR1+Masc.OblSg+Gen+Pl
                            ’rak2+i+bos+kier+es+ke
                                        speak+BND1+OblSg+NR1+Masc.OblSg+Dat
vandrudiijom                ’vandr+uv+di1+l+j2om
                                        wander+INTR1+NFA5  noun+PRET.1SG
vandrula                    ’vandr+ula wander+INTR.IND.PRES.3SG
```

```
rom
’rom2+     Rom+NomSg

1:
     Word
   ____|____
 Word     Ninfl
   |        |
 Stem    N-INFL1
   |       +0
 ROOT    +NomSg
 ’rom2
  Rom
```

```
Word:
[ cat:   Word
  head:    [ animacy:-
             case:  NOM
             number:SG
             pos:   N ]
  root:  Rom
  root_pos:N ]

1 parse found

'rom2+     Rom+AccSg.Inanim

1:
      Word
   _____|______
 Word        Ninfl
   |           |
 Stem       N-INFL1
   |          +0
 ROOT    +AccSg.Inanim
 'rom2
  Rom

Word:
[ cat:   Word
  head:    [ animacy:INANIM
             case:  ACC
             number:SG
             pos:   N ]
  root:  Rom
  root_pos:N ]

1 parse found

rommesta
'rom2+es+ta     Rom+OblSg+Abl

1:
           Word
       ______|_______
     Word       OBLIQUE-SG
   ____|____         +ta
 Word    Ninfl     +Abl
   |       |
 Stem   N-INFL1
   |      +es
 ROOT   +OblSg
 'rom2
  Rom

Word:
[ cat:   Word
  head:    [ animacy:-
             case:  ABL
             number:SG
             pos:   N ]
  root:  Rom
  root_pos:N ]

1 parse found
```

```
romvales
'rom2++val+es     Rom+NomSg+AJR8+Masc.AccSg.Anim

1:
                  Word
          ____________|____________
       Word         Nsuff        Ainfl
     ____|____        |            |
  Word     Ninfl  N-SUFF4      A-INFL1
    |        |     +val          +es
  Stem    N-INFL1  +AJR8 +Masc.AccSg.Anim
    |       +0
  ROOT    +NomSg
  'rom2
   Rom

Word:
[ cat:   Word
  head:    [ aform: ABS
             animacy:ANIM
             case:  ACC
             gender:MASC
             number:SG
             pos:   AJ ]
  root:  Rom
  root_pos:N ]

1 parse found


'rom2++val+es     Rom+NomSg+AJR8+AV1

1:
               Word
          _____|_____
       Word         Nsuff   Ainfl
     ____|____        |       |
  Word     Ninfl  N-SUFF4 A-INFL1
    |        |     +val     +es
  Stem    N-INFL1  +AJR8   +AV1
    |       +0
  ROOT    +NomSg
  'rom2
   Rom

Word:
[ cat:   Word
  head:    [ pos:   AV ]
  root:  Rom
  root_pos:N ]

1 parse found
```

```
tseeresko
'tse1r+es++ko     house, home+AccSg.Anim+CLIT3

1:
              Word
      ____________|____________
     Word         Ninfl    CLITIC
       |            |        ++ko
     Stem        N-INFL3   +CLIT3
       |           +es
     ROOT      +AccSg.Anim
    'tse1r
 house, home

Word:
[ cat:   Word
  head:    [ animacy:ANIM
             case:  ACC
             number:SG
             pos:   N ]
  root:  house, home
  root_pos:N ]

1 parse found

:
:
:
```