HELSINGIN YLIOPISTO
YLEISEN KIELITIETEEN LAITOS
KIELI-, PUHE- JA KÄÄNNÖSTEKNOLOGIAN MAISTERIOHJELMA

Pro gradu -tutkielma

Suomen kielen äärellistilainen morfologinen jäsennin avoimen lähdekoodin resurssein

Tommi Pirinen
013160681

Ohjaaja: Kimmo Koskenniemi ja Krister Lindén 26.4.2008

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos Institution – Department	
Humanistinen tiedekunta	Yleinen kielitiede	
Tekijä Författare – Author		
Tommi Pirinen		
Työn nimi Arbetets titel – Title		
Suomen kielen äärellistilainen morfologinen jäsennin avoimen lähdekoodin resurssein		
Oppiaine Läroämne – Subject		
Kieli-, puhe- ja käännösteknologian maisteriohjelma		
Työn laji Arbetets art – Level	Aika Datum – Month and year	Sivumäärä Sidoantal – Number of pages
Pro gradu – Master's thesis	29.4.2008 – April 2008	64

Tiivistelmä Referat – Abstract

Tutkielma on kuvaus suomen kielen automaattisen äärellistilaisen morfologisen jäsentimen toteutuksesta avoimen lähdekoodin menetelmin ja resurssein. Tutkielmassa kuvataan Kotimaisten kielten tutkimuskeskuksen julkaiseman Nykysuomen sanalistan sisältöä morfologisen jäsentimen toteuttamisessa. Äärellistilaisista menetelmistä tutkielma tarkastelee SFST-nimisen ohjelman käyttämistä jäsennintransduktorin rakentamisessa. Lisäksi tutkielmassa esitetään toteutetun järjestelmän testausmenetelmä valmiin morfologisesti jäsennetyn korpuksen avulla.

Avainsanat – Nyckelord – Keywords
Äärellistilaiset menetelmät, morfologia, suomen kieli

Säilytyspaikka – Förvaringställe – Where deposited
Humanistisen tiedekunnan kirjasto / Yleisen kielitieteen laitos

Muita tietoja – Övriga uppgifter – Additional information

Sisältö

1 Johdanto

Kieliteknologia on monitieteinen ala, joka käsitelee kielitieteellisten teorioiden soveltamista tietojenkäsittelytieteiden menetelmin. Tässä tutkielmassa keskitytään kieliteknologian osa-alueeseen, joka käsittelee sanojen automaattista jäsentymistä ja käsittelyä. Tutkielmassa kuvaan kieliteknologisen suomen kielen sanoja jäsentävän sovelluksen toteutusta äärellistilaisin menetelmin. Sovelluksen erääksi kehitystavoitteeksi on valittu kokonaisjärjestelmän lisenssiehtojen avoimuus ja vapaus. Tutkielmassa kuvaan avointen ja vapaasti saatavilla olevien resurssien käyttöä äärellistilaisen morfologisen jäsentimen rakentamisessa.

Morfologia viittaa tässä tutkielmassa laveasti luonnollisen kielen sanojen muoto-opilliseen rakenteeseen ja sen mallintamiseen. Muoto-opillisen rakenteen osayksiköitä nimitetään konkreettisina instansseina morfeiksi, ja näiden perusteella johdettuina abstrakteina malleina morfeemeiksi. Esimerkiksi jos puhutaan suomen kielen sanamuodosta *taloissa*[1], on siinä juurimorfi *talo*, monikon tunnusmorfi *i* ja inessiivin tunnusmorfi *ssa*. Morfeemeina voisi pitää vaikkapa esitystä *talo*, *I* ja *ssA*, jossa *I* ja *A*[2] kuvaavat abstrakteja morfofoneemeja, jotka toisissa tilanteissa voivat olla vaikkapa *i* ja *a* yhtä hyvin kuin *j* ja *ä* vastaavasti vokaalisoinnun tai vastaavan äännemuunnoksen vaikutuksesta. Tämä luokittelu ja kuvaus vastaa läheisesti sitä, joka on esitetty suomen kielestä mm. teoksessa Karlsson (1982).

Yleensä kielitieteissä sanan jako morfeemeiksi tapahtuu jakamalla sana pienimmiksi semanttista sisältöä kuvastaviksi yksiköiksi (Karlsson, 1998). Jokaisella morfeemilla on oma merkityksensä sanamuodon *taloissa* kokonaismerkityksen kannalta: *talo* merkitsee reaalimaailman talo-tyyppistä oliota, *i* merkitsee monikkoa eli sitä että olioita on monta ja *ssa* merkitsee viittausta talojen sisässä olemiseen. Tutkielmassa kuvatussa järjestelmässä relevanttia on siis se, että *taloissa*, on *talo*-sanan monikon inessiivi. Muihin asioihin kuten sanan rakenteen merkitykseen tai äännerajaukseen lopullinen järjestelmä ei ota kantaa.

Gradun terminologiassa sanan kokoisiin yksiköihin joudun viittamaan monesta eri näkökulmasta, joten olen yrittänyt tehdä tarkan jaon sanan eri merkityksille. Sanalla 'sana'[3] viittaan yhteen sanakirjasanaan sen kaikissa muodoissa, esimer-

[1] Käytän tutkielmassa *kursivointia* viittamaan sanoihin, sanamuotoihin, saneisiin, morfeihin ja morfeemeihin lingvistisinä olioina niiden merkityksen asemesta

[2] Käytän tutkielmassa suuraakkosia kuvaamaan abstrakteja morfofoneemeja. Tavallisesti kuvauksissa vastaavat A, O ja U vokaalipareja a ja ä, o ja ö sekä u ja y, eli sointuvokaalien pareja, koska ne käyttäytyvät lähes poikkeuksetta samoin vokaalisointua lukuunottamatta ja merkintätapa on melko vakiintunut. Vastaavasti merkitään myös V:llä mitä tahansa vokaalia ja C:llä konsonanttia. Tapauksittain myös kirjaimia K, P, T, G, B ja D käytetään kuvaamaan astevaihtelun alaisia klusiileja sekä kirjainta L likvidojen joukkoa

[3] muualla käytetään myös esim. nimitystä lekseemi tai hakusana

kiksi sanaan *paperi*. Yhdellä sanalla voi olla useita sanamuotoja, kuten sanalla *paperi* taivutusmuodot *papereiden* ja *papereitten*. Sanamuodot *papereiden* ja *papereitten* vastaavat samaa morfologista sanamuotoa, monikon genetiiviä, ja niitä kutsutaan tämän muodon allomorfeiksi. Juoksevia tekstejä käsitellessäni jokaista sanamuodon esiintymää kutsun saneeksi.

Tutkielmassa kuvaamani toteutus on **suomen kielen** muoto-opillinen jäsennin, tarkalleen ottaen sen eräs varhainen kehitysversio. Suomen kielen äänne- ja muoto-opin kannalta sanojen taivutuksessa esiintyy monia rinnakkaisia vaihteluja sekä sanavartalossa että tunnuksissa. Esimerkiksi sanan *yö* monikkovartaloksi sanamuodossa *öissä* voisi laskea morfin *ö*, joka poikkeaa yksikön taivutusvartalosta *yö*. Vastaavasti esiintyvän inessiivin tunnuksen *ssä*-muoto johtuu sanavartalon vokaalien etisyydestä, muussa tapauksessa se voisi olla *ssa*. Tällä tarkoitan, että suomen kielen sanojen taivutuksessa esiintyy ainakin sekä juurimorfeemin äännevaihtelua taivutuksen yhteydessä että taivutusmorfeemin äännevaihtelua juuren äännerakenteen vaikutuksesta.

Äärellistilaisilla menetelmillä (finite-state methods) tarkoitan tämän tutkielman piirissä kieliteknologisia sovelluksia, jotka käsittelevät syötteenään merkkijonoja kappaleessa 3.1 tarkemmin kuvatun mekaniikan mukaisesti. Äärellistilaisista olioista puhun lähinnä *transduktorista* (transducer), joka on käytännön tasolla sellainen mekanismi, joka ottaa syötteekseen merkkijonon, tunnistaa sen ja muuntaa toisiksi vastineiksi. Äärellinen transduktori on nimenomaan morfologian kannalta käyttökelpoinen, sillä niillä voi kuvata sanan sanakirjamuotojen ja taivutusmuotojen välisen suhteen suoraan, ja tämä kuvaus on valmiiksi kaksisuuntainen siten, että se osaa muuntaa sanan ja taivutusmuotojen tunnisteet sanamuodoksi sekä päinvastoin. Käytännössä siis eräs yksinkertainen transduktori olisi sellainen, jossa on kuvaus `talo SG INE` $\leftrightarrow$ `talossa`[4]. Tässä tutkielmassa esitellään transduktori, joka kuvaa tutkielman sanalähteenä olevan sanalistan kaikki sanat kaikille suomen kielen säännöllisen taivutusopin sanamuodoille.

Avoin lähdekoodi (open source) on valittu keskeiseksi perusteeksi tutkielmassa käytettävien aineistojen lisensointiin, sillä on pidetty ongelmana, että saatavilla olevat suositut resurssit, kuten äärellistilaisia järjestelmiä käsittelevät koodit tai sanakirjalähteet, ovat tarkoin määritellyin lisenssein kaupallisia tai suljettuja ohjelmia. Tämä hankaloittaa niiden yleistä kehitettävyyttä sekä niiden avulla käsitellyn materiaalin vapaata jakelua (ks. esim. Yli-Jyrä (2005)). Avoimella lähdekoodilla tässä viittaan ensi sijassa sellaisiin resursseihin, joiden sisältö on vapaasti saatavilla, mukautettavissa ja edelleenlevitettävissä lisenssinsä perusteella.

Järjestelmän toteuttamiseen tarvittavia äärellistilaisia ohjelmistoja on saatavilla

[4]Merkitsen tutkielmassa koodijärjestelmien merkintöjä, kuten tunnisteita, nimiä ja varattuja sanoja `tasavälisellä fontilla`.

runsaasti. Kuvaan tässä lyhyesti ominaisuuksien valintaa ja niiden relevanssia tutkielman sekä siinä toteutetun järjestelmän kannalta, minkä jälkeen tarkennan asiaa valitun järjestelmän kannalta. Liitteessä A on taulukko, joka sisältää tietoja kaikista kirjoitushetkellä löydetyistä ohjelmistoista (Yli-Jyrä, 2007), sikäli kun ne ovat löydettävissä, tarkastettavissa tai muuten tunnetut.

Koska tutkielman tavoitteena oli alusta loppuun rakentaa avoin ja vapaa kokonaisjärjestelmä, ohjelmistojen lisenssiehdoista on otettu huomioon ensi sijassa avoimuus. Tällä perusteella myös voi sinänsä sivuuttaa kaikki suljetut ohjelmistot irrelevantteina. Tyypillisimpiä avoimen lähdekoodin lisenssejä ovat GNU-projektin (GNU is Not Unix) GPL- (GNU General Public Licence) ja LGPL-lisenssit (GNU Lesser General Public Licence), joita käsiteltäessä on pantava merkille avoimuuden lisäksi muitakin rajoituksia tai mahdollisuuksia[5]. GNU-lisenssejä pidän ensisijaisina avoimina ja vapaina lisensseinä tässä selvityksessä.

Sovellusten toiminnallisista ominaisuuksista on valitussa järjestelmässä pidetty tärkeänä toisinkirjoitus- eli replace-sääntöjen kirjoittamisen mahdollisuutta, ja replace-sääntöjä on käytetty toteutuksen perustana. Replace-säännöstä kerron tarkemmin kappaleessa 3.1. Käytännössä kyseessä on operaatio, jolla voi yksinkertaisesti ja havainnollisesti mallintaa yleisiä suomenkin kielen morfofonologisia vaihteluita. Esimerkiksi muutamalla tällaisella säännöllä voi mallintaa taivutuspäätteen sointuvokaalin valintaa suhteessa sitä edeltävään lähimpään sointuvokaaliin. Ne mahdollistavat tavan ilmaista helpohkosti, että jos vasemmasta kontekstista löytyy takainen sointuvokaali ennen lähintä sananrajaa, käytetään takaista sointua, muutoin etistä. Yksi tällainen sääntö voisi olla yksinkertaistaen vaikkapa `A:ä → ([äöy] __)`: sointuvokaalista tulee ä jos vasemalla on ä, ö tai y.

Tutkielmassa toteutustyökaluksi on valittu SFST-ohjelmisto, joka on GNU GPL -lisensoitu ja sisältää replace-säännön sellaisen toteutuksen, joka on suomenkielisen morfologisen jäsentimen toteuttamiseen riittävä. SFST:tä ei aiemmin ole käytetty kuin muutamissa saksan kielen taivutuksen ja johto-opin morfologiaa kuvaavissa järjestelmissä (Schmid, 2005), mutta hyötypuolena järjestelmää ja toteutusta verrattaessa muihin saatavilla oleviin, oli mukana edes jonkinlaiset ohjeet ja kuvaukset morfologian toteuttamisesta (Schmid, 2007a,b). Joitakin sekä lisenssiltään että toiminnoiltaan soveltuvia ohjelmistoja, kuten PC-KIMMOa KGENillä tai mmorphia ei ole tarkasteltu, sillä niiden saaminen toimimaan modernilla kääntäjillä ja järjestelmäkokoonpanoilla osoittautui hankalaksi tai mahdottomaksi ilman merkittäviä muutoksia koodiin. Mitä tulee OpenFST:hen, joka on mm. AT&T:n FSM:n tekijöiden julkaisema avoin ja vapaa FST-toteutus, tiedon siitä sain vasta syksyllä 2007, jolloin oma morfologiani SFST:llä oli jo lähes valmis, ja

[5]GNU-lisensseistä ja käytöstä sekä periaatteista tarkemmin ks. `http://www.gnu.org/philosophy/philosophy.html`.

se on jätetty siitä syystä pois laskuista. Jatkotutkimusta varten kuitenkin OpenFST lienee varsin kehityskelpoinen aihe.

Vertailukohdaksi suorituskykytestauksessa on valittu kaupallinen ja tutkimuskäyttöön rajattu AT&T:n FSM-järjestelmä, joka lienee ainakin tunnettu ja käytetty. Se oli myös saatavilla suorituskykytestauksia varten, ja muunnos käyttämästäni järjestelmästä sille sopivaksi testaustarkoituksiin ei ollut kovin hankala.

Toinen merkittävä resurssi, jonka tässä automaattisen morfologisen jäsentimen toteutuksessa tarvitsin, oli leksikko. Tätä tarkoitusta varten suomenkielinen sanalista, Kotimaisten kielten tutkimuskeskuksen GNU LGPL -lisenssillä julkaisema Nykysuomen sanalista (Kotimaisten Kielten Tutkimuskeskus, 2006), oli juuri julkaistu tutkielman teon vaiheessa ja sopi loistavasti tutkittavaksi materiaaliksi. Nykysuomen sanalistan muoto, joka on seikkaperäisemmin kuvattu kappaleessa 2, määritteli myös luontevasti rajat ja tyylin tekemälleni järjestelmälle, sillä siinä sanat on luokiltaan jaettu perinteisen sanakirjaluokituksen mukaisesti, joka jotakuinkin kuvaa morfofonologisia piirteitä sanoissa, ja jota olen jokseenkin tarkasti seurannut järjestelmässäni. Toteutus toimii siis sillä perusolettamuksella, että tietyllä numerolla olevassa taivutusluokassa sana taipuu aina saman kaavan mukaisesti, ja tällainen kaava väistämättä pätee kaikkiin luokan sanoihin. Sanalistassa ei toisaalta ole eroteltu esimerkiksi adjektiiveja substantiiveista, jonka perusteella rajoitin myös adjektiivitaivutuksen järjestelmän ulkopuolelle. Toteutin myös testaustarkoituksiin yligeneroivan yhdyssanamuodostuksen, joka yhdistelee taivutettuja nomineja melko mielivaltaisesti. Sanojen johtamista ei myöskään ole toteutettu verbien yleisempiä infiniittimuotoja enempää.

Kolmantena resurssina järjestelmän testaukseen tarvitaan saneita oikeine morfologisine tulkintoineen. Tällaisia valmiiksi jäsennettyjä tekstiaineistoja kutsutaan myös korpuksiksi. Sopivia korpuksia ei ole vapain lisenssein, joten tutkin tutkielmassani rajoitetummin lisensoidun automaattisesti jäsennetyn tarkistamattoman korpusmateriaalin käyttöä kertaluontoisen testauksen toteuttamisessa. Lisensoinnin takia testaus jouduttiin suorittamaan sivullisella palvelimella eikä tuloksia ole mahdollista hyödyntää vertailua laajemmin niiden lisenssiehtojen puitteissa.

Kokonaisuutena graduni on selvitystyö, johon yritin koota tietoa seuraavista:

1. morfologiajärjestelmän luonnosteleminen, jos saatavilla on jo luokiteltu sanalista, sekä tunnettu säännöllinen taivutusjärjestelmä
2. Nykysuomen sanalistan datan käyttö koneellisen suomen kielen morfologian toteutukseen
3. Nykysuomen sanalistan taivutusluokkien käyttö koneellisessa morfologiajärjestelmässä

4. Nykysuomen sanalistassa käytetyn XML-formaatin käyttö morfologisen järjestelmän osana
5. SFST:n käyttö suomen kielen morfologisen järjestelmän teossa
6. SFST:n suorituskyky verrattuna AT&T:n FSMlib-järjestelmään
7. morfologisen järjestelmän testaaminen korpusaineistolla, joka on morfologisesti jäsennetty.

Tutkielma jakautuu selvitykseltään kahteen pääteemaan. Ensin käydään läpi käytössä olevien resurssien ja menetelmien sisältö ja käyttötarkoitukset. Toiseksi tutkitaan näin resurssein valmistamani järjestelmän toimintaa testaamalla sitä erilaisin aineistoin ja suorituskykytestein ja analysoimalla havaitut puutteet ja ongelmat. Ensimmäinen osa rakentuu tutkielmassani siten, että kappaleessa 2 selvitetään Nykysuomen sanalistan sisältöä siltä kannalta, miten sitä käytetään morfologisessa jäsentimessä. Selvityksessä käydään läpi yleinen morfofonologinen tieto suomen kielen taivutuksesta ja se, miten se suhtautuu sanalistassa olevaan luokitukseen. Tämän jälkeen kappaleessa 3 kerrotaan äärellistilaisista menetelmistä morfologisen jäsentämisen käytössä. Kappaleessa selvitetään miten kappaleessa 2 selvitetty tieto sanalistan taivutusluokituksesta on muunnettavissa äärellistilaisen järjestelmän käyttöön. Toinen osa koostuu testiaineistojen ja -asetelmien kuvauksista sekä tuloksista kappaleessa 4, jonka jälkeen kappaleessa 5 analysoidaan koko järjestelmää ja sen testituloksia sekä tältä pohjalta asetetaan mahdollisia tulevaisuuden kehityslinjoja ja tutkimusideoita.

2 Aineistot

Automaattinen morfologinen jäsennin, jollaista tutkielmassa rakennetaan, perustuu sanalistaan. Tällaiseksi sanalistaksi on valittu Kotimaisten kielten tutkimuskeskuksen, eli Kotuksen, julkaisema Nykysuomen sanalista, jonka tutkielmassa käytetty julkaistu ensiversio sisältää 94 110 sanaa. Sanoista 44 348:aan on merkitty tieto taivutusluokista sekä mahdollisesta astevaihtelusta, jota sana suomen kielen perussanakirjan taivutusluokittelussa käyttää (Kotimaisten Kielten Tutkimuskeskus, 2006). Loput luokittelemattomat sanat ovat yhdyssanoja, joiden taipuva osa on luokiteltu sanalistassa; nämä sanat olen jättänyt tutkielmassani toistaiseksi huomiotta.

Nykysuomen sanalista on tallennettu XML 1.0[6] -määrittelyn mukaisen merkkauskielen sovelluksella. XML-kielen käyttö tarkoittaa että tiedosto on pohjimmiltaan

[6] `http://www.w3.org/TR/2006/REC-xml-20060816/`

tekstipohjainen puuta kuvaava datarakenne, jonka osaset on merkattu kulmasulkeisiin rajatuin tunnistein eli merkkauksin. Sanalistassa se tarkoittaa pelkistäen, että jokaista sanaa kohti on eroteltu vähintään tiedot sanan sanakirjamuodosta, taivutusluokasta ja mahdollisesta astevaihteluluokasta. Tyypillinen sanatietue sisältää siis osat `<s>sanakirjamuoto</s>`, `<tn>taivutusluokka</tn>` ja `<av>astevaihtelukirjain</av>`. Kokonaisia esimerkkejä on listauksessa 1.

Ensimmäinen toteutuksen kannalta olennainen asia Nykysuomen sanalistan muuttamisessa kohti morfologista jäsennintä on selvittää mitä tietoja sanalistasta jäsennin tarvitsee voidakseen kuvata sanakirjamuodot taivutusmuodoiksi. Lähtökohta toteutuksessani on, että sanalistassa käytetty sanaluokitus on itsenäisesti riittävä kuvaamaan kaikki tai lähes kaikki morfofonologinen vaihtelu, joka sanojen taivutuksen kuvauksessa on tarpeellista. Seuraavaksi selvitetäänkin sanalistassa käytetyn luokituksen periaatteet ja miten ne muuntuvat äärellistilaisen jäsentimen käytettäväksi.

Nykysuomen sanalistan data — sanat ja niiden luokitukset — suurimmilta osin soveltuvat sellaisenaan lopullisen järjestelmän käyttöön, eikä niitä tarvinnut jakaa, yhdistellä tai muuttaa, joten kappaleessa 2.1 kuvaan tämän alkuperäisen XML-rakenteen osineen. Muutamat muutokset ja päivitykset, jotka olen tehnyt sanalistaan on myös selostettu. Tämän jälkeen kappaleessa 2.2 kuvaan mitä sanalistan sisältämä data tarkoittaa käytännössä lingvistiseltä kannalta, eli mitä kaikkea sellaista käytetyn sanaluokituksen luokat kuvaavat, mikä morfologisen jäsentimen tulee ottaa huomioon.

2.1 Nykysuomen sanalistan XML-muoto

Nykysuomen sanalista on yksinkertainen XML-muotoinen tiedosto. Se sisältää yhden monialkioisen listan, jonka jokaisen alkion sisältönä on datarakenne sanatietue-alkiossa `st`, johon kuuluu sana sanakirjamuodossaan alkiona nimeltä `s`. Toinen osa sanan sisältävää datarakennetta on taivutustiedot alkiossa `t`, johon kuuluu sanan taivutusluokka alkiona `tn` sekä mahdollinen astevaihtelutieto alkiona `av`. Taivutustietoja per sana voi olla useampia kuin yksi, mutta vain erikoistapauksissa (harvinainen vanhempi taivutus tms., jolloin `t`-alkiolla myös on selittävä vakioitu tekstimuotoinen attribuutti), sillä homonyymit on merkitty eri sanoiksi, ja niille on datarakenteessa erillinen alkio `hn`, joka on juokseva homonyyminumero.

Kuvasta 1 näemme esimerkkejä miten erilaiset sanatyypit kuvataan tässä XML-muodossa. Kuvaan tässä XML-rakenteeseen kuuluvat alkiot siinä järjestyksessä, jossa ne tulevat esille. Juuren muodostaa XML-alkio `kotus-sanalista`, jolla

Kuva 1: Otos Nykysuomen sanalistan XML-datasta (Kotimaisten Kielten Tutkimuskeskus, 2006)

```
<st><s>aloitteikas</s><t><tn>41</tn><av>A</av></t></st>
<st><s>-aloitteinen</s><t><tn>38</tn></t></st>
<st><s>aloittelija</s><t><tn>12</tn></t></st>
<st><s>aloitus</s><t><tn>39</tn></t></st>
<st><s>aloituskorkeus</s></st>
<st><s>aloitusmerkki</s></st>
<st><s>aloituspaikka</s></st>
<st><s>aloitussyöttö</s></st>
<st><s>aloitusviisikko</s></st>
<st><s>alokas</s><t><tn>41</tn><av>A</av></t></st>
<st><s>alokasaika</s><t><tn>9</tn><av>D</av></t></st>
<st><s>alokasaste</s></st>
<st><s>alokasmainen</s><t><tn>38</tn></t></st>
<st><s>aloke</s><t><tn>48</tn><av>A</av></t></st>
<st><s>alpakka</s><hn>1</hn>
  <t><tn>14</tn><av>A</av></t></st>
<st><s>alpakka</s><hn>2</hn>
  <t><tn>14</tn><av>A</av></t></st>
<st><s>alpakkainen</s><hn>1</hn><t><tn>38</tn></t></st>
<st><s>alpakkainen</s><hn>2</hn><t><tn>38</tn></t></st>
<st><s>alpakkalusikka</s></st>
<st><s>alpi</s><t><tn>7</tn><av>E</av></t>
  <t taivutus="harvinainen"><tn>5</tn></t></st>
```

ei ole attribuutteja, ja sisältönä on nolla tai useampia sanatietue-alkioita, eli se on yksinkertainen listan sisällyttävä rakenne.

Sanatietueita ovat `st`-alkiot, eikä niilläkään ole mahdollisia attribuutteja. Sisältönään jokaiseen sanatietueeseen kuuluu tasan yksi sana-alkio, nolla tai yksi homonyymialkiota ja miten monta tahansa taivutustietoalkiota.

Sana-alkiolla, jonka nimi on `s`, ei ole mahdollisia attribuutteja, ja sen sisältönä on pelkkää tekstiä. Tekstisisällöksi kuuluu sana sanakirjamuodossaan. Tämä tarkoittaa tyypillisesti nomineilla yksikön nominatiivia ja verbeillä a-infinitiivin latiivia. Poikkeuksia muodostovat esim. nomineista monikkosanat, joiden sanakirjamuotona käytetään monikon nominatiivia (esim. *häät*, *sakset*), ja verbeistä sanat, joilla a-infinitiivin latiivia ei käytetä (esim. *erkanee* pro $^{?}$*erata*).

Homonyymialkiolla `hn` ei ole attribuutteja. Sen sisältönä on positiivinen kokonaislukuarvo, joka ei saa olla sama kuin yhdelläkään toisella saman sanan (so. sellaisen, jonka `s`-alkion sisältö on sama) sanatietueella. Käytännössä siis sanalistassa homograafit numeroidaan juoksevalla tunnisteluvulla.

Taivutusalkiolla `t` on valinnainen attribuutti `@taivutus`, jonka arvoina on sanallinen selitys taivutusalkion kuvaaman taivutuksen poikkeusluonteesta. Nykysuomen sanalistan versiossa 1 attribuutissa käytetyt arvot olivat `"harvinainen"` ja `"mahdollinen"` kuvaamaan vaihtoehtoisten taivutusten yleisyyttä sekä `"yksikössä"` ja `"monikossa"`, jolla kuvattiin sanan *kolme* taivutustapaa. Omassa koeversiossani laajensin attribuuttiarvoa `monikossa` koskemaan monikkosanojen `t`-alkiota. Taivutusalkion sisältönä on tasan yksi taivutusnumeroalkio, sekä nollasta yhteen astevaihtelualkioita.

Taivutusnumeroalkio `tn`:llä ei ole attribuutteja, ja sen sisältönä on positiivinen kokonaisluku, joka on yksi kappaleessa 2.2 kuvatuista taivutusluokista.

Astevaihtelualkiolla `av` on attribuutti `astevaihtelu`, jonka ainoa arvo sanalistassa on `valinnainen`, niille sanoille joita voi taivuttaa sekä astevaihtelullisena että -vaihteluttomana. Alkion sisältönä on suuraakkonen alueelta A—M, joka kertoo astevaihtelun alaisena olevan äänteen ja sen heikon asteen vastineen.

Taivutusnumero ja astevaihtelukirjain yhdessä sanan sanakirjamuodon kanssa muodostavat sen keskeisen datan, johon nojaten olen jäsentimeni rakentanut.

Näiden lisäksi olen muokannut hieman formaattia siten, että on mahdollista kuvata sanan sanakirjamuodon on monikollisuus (esim. *sakset* tai muut pluratiivit), sanan vokaalisoinnun tai vartalovokaalin odotuksenvastaisuus (esim. *bordeaux* ja monet vierassanat), ja sanan taivutuksen odotuksenvastaisuus (esim. *olla*). Näistä ja muista kehitysideoista tarkemmin 5. kappaleessa.

2.2 Nykysuomen sanalistan sisältö: sanaluokitus käytännössä

Tässä kappaleessa pyrin selvittämään Nykysuomen sanalistassa käytetystä taivutusluokituksesta kaiken sen tiedon, jota lopullisessa morfologisessa jäsentimessä on käytetty. Kyseessä on käytännöllinen kuvaus siitä, miten taivutusluokittelua lopulta päädyin käsittelemään järjestelmässä, ei niinkään lingvistinen kuvaus taivutusluokituksen sisällöstä. Paikoin kuvauksissa on annettu viitteitä relevantteihin lähteisiin suomen kielen äänne- ja muotohistoriasta, mutta lähinnä niiltä osin kuin annettu lisätieto riittää motivoimaan järjestelmässä esiintyvän äännevaihtelun. Systemaattiset selvitykset luokituksesta kattavat vain näkyvissä olevat konkreettiset muutokset sanojen kirjoitusmuodossa sikäli kuin ne tämän järjestelmän toteutuksessa ovat olleet tarpeellisia.

Nykysuomen sanalistan sanat on luokiteltu saman luokittelun mukaisesti kuin mitä on käytetty Suomen kielen perussanakirjassa ja Kielitoimiston sanakirjassa. Nomineille tarkoitettuja luokkia on 49 (1—49) ja verbeille 27 (52—78). Lisäksi sanalistassa on yksi luokka kaikille adverbeille, adpositioille, lyhenteille ja partikkeleille (99), yksi pronomineille (101) ja kaksi yhdysnomineille (50 ja 51), joita ei tämän tutkielman morfologian toteutuksessa lainkaan käsitellä. Sanaluokitus perustuu vanhempaan Nykysuomen sanakirjan luokitukseen siten, että luokkien määrää on vähennetty fonologisin perustein (Eronen, 1994).

Sanaluokitus vaikuttaa enimmäkseen käytännössä toteutetun siten, että numerosta käy ilmi sekä sanalle kuuluvat potentiaaliset allomorfit että vartalossa esiintyvät äännevaihtelut, toisin sanoen jos yhdenlaisista perusmuodon vartalon äännevaihtelusta koostuvalla luokalla esiintyy jonkin taivutuspäätteen kohdalla erilaista allomorfien jakaumaa sen pitäisi aina saada oma taivutusluokkansa. Poikkeuksena päätevaihtelun tuomista taivutusluokista on vokaalisoinnun aiheuttamat muutokset, joita ei ole merkitty taivutusluokiksi, ja joka pitää tässä huomioida morfologian toteutuksessa sääntöpohjaisesti. Vastaavasti vartalon äännemuutoksista astevaihtelu ei tuo uusia taivutusluokkanumeroita, sillä se on kuvattu erillisenä komponenttina taivutusluokitusta astevaihtelukirjaimella. Tästä seuraten siis sanat *hyytö* ja *huuto* ovat molemmat luokassa 1, vaikka yksikön inessiivin tunnukset *ssä* ja *ssa* poikkeavat, kun taas sanat *valo* ja *valtio* taas luovat eri luokat 1 ja 2, koska monikon genetiivin tunnusten *jen* ja *iden* poikkeukset eivät selity vokaaliharmoniasäännöllä.

Taivutusluokitus on toteutetussa järjestelmässä analysoitu eli eroteltu piirteiksi niin, että taivutusluokkanumeron oletetaan kuvaavan:

- taivutuksessa vartaloon kohdistuvat **äännemuutokset**
- taivutustunnusten **allomorfien** valinnat niissä taivutuspäätteissä, joissa

esiintyy taivutusluokittaista vaihtelua

- astevaihtelun tyypin (suora vai käänteinen astevaihtelu)
- perusmuodon **äännerakenteen**.

Astevaihtelukirjain taas kuvaa suoraan astevaihtelun alaisen äänneparin, josta yhdessä astevaihteluluokan kanssa on selvitettävissä kirjain, joka vastaa astevaihtelun käyvää äännettä sanan sanakirjamuodon kirjoitusasussa. Astevaihtelun kuvauksesta erillisenä yhdistelmällisenä piirteenä luokituksessa seuraa myös, että astevaihtelun vaikutusta äännerakenteeseen ja vartalon äännemuutoksiin ei oteta lukuun niitä kuvatessa.

Seuraavassa käyn läpi taivutusluokkia yksi kerrallaan siten, että selvitän jokaisesta perusmuodon äännerakenteesta luokalle yhteisen osan, joka tässä tarkoittaa sanan perusmuodon loppua, jokaisen allomorfista vaihtelua sisältävän päätetyypin sallitut allomorfit, sekä taivutuksessa sanan vartalossa esiintyvät äännemuutokset. Näitä ei ole tietääkseni missään erikseen kuvattu tämän taivutusluokituksen suhteen, joten kuvaukseni perustuu toisaalta tietoihini suomen kielen äänne- ja muotorakenteen kehityksestä, joista lähteinä järjestelmää työstäessäni olen käyttänyt alkeisteoksia Remes (2004); Setälä (1930); Karlsson (1982), toisaalta puhtaasti kokeiluun ja kielitajuuni. Alkeisteoksista Setälän kielioppi on yksi selkeimmistä ja systemaattisimmista näkemistäni esityksistä suomen sanojen muoto-opin järjestelmästä, vaikkakin paikoin vanhentunut, joten olen käyttänyt myös Remeksen laatimaa luentomonistepakettia faktojen tarkistuksessa. Samat tiedot löytynevät fennistiikan peruskirjallisuudesta muutoinkin. Karlssonin äänne- ja muoto-oppia taas on käytetty Nykysuomen sanakirjan luokkien vähentämiseksi Suomen kielen perussanakirja, joten olen siitä poiminut myös joitain kuvauksia[7].

Nykysuomen sanalistan luokittelu lienee motivoitu siten, että sanaluokkien ensimmäiset luokat 1 ja 52 ovat perusluokkia, joissa äännevaihtelu on vähäisintä, ja uudet luokat on tuotu jokaiselle havaitulle uudelle äännevaihtelukokonaisuudelle tai pääteallomorfijoukolle. Samalla perusteella itse tutkielman toteutettu morfologiakin on kappaleessa 3 kuvattu. Kaikkien taivutusluokkien erittely löytyy myös kuvausten jäljestä taulukoista 1, 2 ja 3.

Nominit jakautuvat tavallaan kahteen pääluokkaan: ne, joiden perusmuoto on vokaalivartalo, ja joiden astevaihtelu on suora, muodostavat luokat 1—32, ja ne, joiden perusmuoto on konsonanttivartalo, ja joiden astevaihtelu on käänteinen, muodostavat luokat 33—49. Verbit jakautuvat samoin astevaihtelun suhteen suoraan

[7]Tiedon Karlssonin kirjan käytöstä sanakirjatyössä sain vasta laadittuani seuraavat kuvaukset ja taulukot, joten sitä olen käyttänyt viitteenä jälkikäteisesti.

(luokat 52—65) ja käänteiseen (luokat 66—78). Verbeistä lisäksi huomionarvoista on, että sanakirjamuoto a-infinitiivissä on taivutusvartalon lisäksi A-, dA- tai tA-tyyppinen a-infinitiivin tunnus.

Ensimmäinen taivutusluokka (*valo*) on nominien yksinkertaisin luokka, johon ei liity ollenkaan vartalon äännemuutoksia poislukien sanaluokkakoodiin erikseen merkityt astevaihtelumuutokset, ja vaihtelevia allomorfeja ei ole vaan monikon genetiivin, partitiivin ja illatiivin tunnuksissa on vain yksi kieliopillinen vaihtoehto, morfeemit *jen* (*valojen*), *jA* (*valoja*) ja *ihin* (*valoihin*) vastaavasti. Luokkaan kuuluvien nominien vartalon äännerakennetta rajoittaa vain vartalovokaalina O tai U, jotka ovat vaihteluttomia. Luokka on suurehko ja avoin, sillä siihen on merkitty muun muassa nut-partisiipin passiivin[8] kiteytyneet muodot sekä stO-johdokset, jotka molemmat ovat hyvin produktiivisia[9].

Luokat 2—4 ovat muutoin kun luokka 1, mutta ne merkitsevät erilaisia päätteallomorfivaihtoehtoja.

Luokan 2 (*palvelu*) päätteiden tunnusten allomorfeihin kuuluu luokan 1 morfeemien lisäksi monikon partitiivin tunnus *itA* (*palveluita*), sekä genetiivin tunnukset *iden* (*palveluiden*) ja *itten* (*palveluitten*). Luokkaan kuuluvien sanojen vartalo on poikkeuksetta vähintään kolmitavuinen, ja lisäksi *O*- tai *U*-loppuinen kuten luokassa 1.

Luokka 3 (*valtio*) on äännerakenteeltaan monitavuisten vokaaliyhtymäpäätteisten sanojen luokka, joiden monikon genetiivin tunnuksen allomorfit ovat *iden-itten*-tyyppiä (*valtioiden* ∼ *valtioitten*), mutta partitiivin vain *itA*-tyyppiä (*valtioita*).

Luokkaan 4 (*laatikko*) kuuluu useita monikon tunnusten allomorfeja, jotka liittyvät sekä heikkoon että vahvaan vokaalivartaloon siten, että muodot *jen* (*laatikkojen*) ja *jA* (*laatikkoja*) liittyvät vahvaan, *iden* (*laatikoiden*), *itten* (*laatikoitten*)avonen] by X-bar 10:38 -!- theriel [theriel@ ja *ita* (*laatikoita*) heikkoon ja *ihin* kumpaankin vartaloon (*laatikoihin* ∼ *laatikkoihin*), kuten on morfofonotaktisesti tyypillistä. Tämä vaihtelu esiintyy vain äännerakenteeltaan kolmi- tai useampitavuisissa sanoissa joiden loppu on *kkO*, kuten tässä luokassa, tai *kkA*, kuten luokassa 14.

Luokat 5 ja 6 (*risti* ja *paperi*) sisältävät vartalovokaalin *i* vaihtelusta *e*:ksi mo-

[8]Käytän tässä Ison suomen kieliopin (Hakulinen et al., 2004) termejä, sillä erotuksella, että kirjoitan tunnusten nimet pienaakkosin, eli nut- pro NUT-partisiippi. Vanhastaan kieliopeissa käytetty muoto on I partisiippi tai menneen ajan partisiippi

[9]Tässä tutkielmassa produktiivisuuden määritelmänä käytetään vain natiivin kielenpuhujan intuitiota morfeemin esiintymän tuoreudesta. Todellisuudessa produktiivisuuden määrittäminen on hankalampi kysymys, jota on tutkittu jonkin verran. Tutkielman kannalta tärkeää on huomata, että produktiivisiksi merkityt morfologiset prosessit ovat sellaisia, joita esiintyy paljon, ja epäproduktiiviset sellaisia, joita ei ole löydettävissä kuin muutama laajoistakin aineistoista.

nikon tunnuksen *i*:n edeltä, eli niihin kuuluu *i*:hin loppuvia sanoja, joiden monikkomuodoissa monikon tunnuksen *i*:tä tai *j*:tä edeltää useimmissa muodoissa *e*. Kuitenkin niin että monikon nominatiivilla tai akkusatiivilla on yksikkövartalon mukainen *i*.

Luokalla 5 monikon genetiivi on *ien*-tyyppiä (*ristien*) ja partitiivi *jA*-tyyppiä (*ristejä*). **Luokkaan 6** kuuluvat sanat saavat useampia monikon tunnusten allomorfeja (*papereiden : papereita*).

Luokka 7 (*ovi*) sisältää myös vaihtelua *i : e*, tai tarkemmin *i : e : ∅*, tässä tapauksessa *i* reaalistuu vain yksikön nominatiivissa ja muissa yksikkömuodoissa esiintyy *e* (*ovena*), joka teoriassa on vartalovokaali. Monikon tunnuksen *i*:n edeltä (*ovina*) vartalovokaali katoaa.

Luokka 8 (*nalle*) on sama vaihteluton taivutusluokka kuin 1, sillä poikkeuksella että siihen kuuluu tyypillisesti vaihteleva vartalovokaali *e*, joka ei tässä vaihtele eikä aiheuta astevaihtelua. Luokkaan kuuluvat sanat ovat lähinnä verrattain uusia, ja monet myös sitaattilainoja, joiden taivutus on muutoinkin epäsäännöllistä, esimerkiksi astevaihtelutonta pääsäännön vastaisesti. Nykysuomen sanalistan mallitaivutustaulukko merkitsee tähän taivutusluokkaan myös harvinaiseksi monikon genetiivin *in*-tunnuksen (*nallein*), joka esiintynee pelkästään tyylivärisenä vanhastavana muotona.

Numeraali *kolme* on merkitty taivutettavaksi sekä luokan 8 että luokan 9 kuvauksen mukaisesti riippuen siitä taivutetaanko yksikössä vai monikossa.

Luokat 9—15 ovat vartalovokaalin *a* vaihteluiden *a : ∅* ja *a : o : ∅* sekä päätealloomorfivarianttien eri kombinaatioita.

Luokassa 9 (*kala*) vartalovokaali *A* reaalistuu *O*:na monikon tunnuksen *i*:n (*kaloissa*) edellä. Monikon taivutustunnuksilla ei ole vaihtelevia allomorfeja.

Luokkaan 9 on sanalistassa merkitty myös sana *aika* (yhdyssanamuotoineen), joka ei siihen säännölisesti sovellu, vaan sisältää myös *i : j* -vaihtelun astevaihtelussa kadonneen koon takia vokaaliväliin joutuneen *i*:n asteittaisen muuttumisen seurauksena. Lisäksi luokassa on numeraali *sata*, joka taipuu kuten odotettua.

Luokka 10 (*koira*) on vartalovokaalin vaihtelun *A : ∅* luokka, jossa monikon tunnuksen edellä *A*:ta vastaa kato. Luokkaan 10 kuuluu myös sana *poika*, jossa esiintyy luokan muista sanoista poiketen yllämainittu *i : j* -vaihtelu. Lisäksi luokkaan on merkitty numeraalit *neljä* ja *nolla*, jotka taipuvat luokan kuvaamalla tavalla, mutta myös numeraalit *seitsemän*, *kahdeksan* ja *yhdeksän*, joiden taivuttaminen luokan kuvaamalla tavalla olettaa, että sanan taivutusvartalona käytetään muotoa, josta loppu-*n* on poistettu.

Luokkaan 11 (*omena*) sisältää vartalovokaalin vaihtelusarjan *A : O : ∅* ja siihen

liittyy koko joukko monikon tunnusten allomorfeja. Monikon genetiivissä tunnukset ovat *ien* (*omenien*), *iden* (*omenoiden*) ja *itten* (*omenoitten*) sekä harvinaisina *jen* (*omenojen*) ja *in* (*omenain*), joista ensi mainittu liittyy vokaalittomaan vartaloon ja jäljemmät *O*:lliseen, sekä harvakäyttöinen *in* yksikkövartaloon. Vastaavasti monikon partitiivilla vokaalittomaan vartaloon liittyy allomorfi *iA* ja *O*:lliseen *itA* sekä harvemmin *jA*. Myös illatiivilla on tässä luokassa kaksi allomorfia, *iin* vokaalittomaan ja *ihin O*:lliseen vartaloon.

Luokkaan 12 (*kulkija*) pätee vain *A : O* -vaihtelu ja ainoastaan monikon genetiivi saa allomorfit *iden* (*kulkijoiden*) ja *itten* (*kulkijoitten*), sekä harvinaisena yksikkövartaloon in-muodon.

Luokka 13 (*katiska*) on kuten 12, mutta allomorfivalikoimaan lisätään monikko genetiivin *jen*-tunnus (*katiskojen*) ja partitiivin *jA*-tunnus (*katiskoja*), kun 12:ssa oli vain *itA*.

Luokka 14 (*solakka*) on kuten 13, mutta siihen liittyy sama monikon illatiivin asteen vapaa vaihtelu kuin luokkaan 4 (*solakoihin* $\sim$ *solakkoihin*). Äännerakenteltaan luokka sisältää vähintään kolmitavuisia sanoja, jotka päättyvät enimmikseen *kkA*, mutta myös *ttA* ja *ppA*.

Luokka 15 (*korkea*) sisältää *A*-loppuisten vokaaliyhtymäloppuisten adjektiivien *a : ∅* -vaihtelun.

Luokkaan 16 (*vanhempi*) on kerätty adjektiivien komparatiivien kiteytymiä, joiden lopun *mpi*-aines taipuu normaalien adjektiivien komparatiivien tavoin *mPA*, ja monikon allomorfit *mp*-vartalon perään tulevat partitiivin *iA*, illatiivin *iin* ja genetiivin *ien*. Myös komparatiiviperäiset pronominit, kuten *kumpi* ja *kumpikin* on sijoitettu tähän luokkaan.

Luokat 17 ja 18 sisältävät pitkän vartalovokaalin vaihtelun lyhyeksi monikon tunnuksen *iin* edeltä. Lisäksi 17 (*vapaa*) ottaa yksikön illatiivin allomorfin seen (*vapaaseen*), joka liittyy vain pitkän vokaaliaineksen perään, ja 18 (*maa*) allomorfin *hVn* (*maahan*).

Luokka 19 (*suo*) on yksitavuisten sanojen väljenevien diftongien suppea ja ilmeisesti suljettu luokka, johon kuuluu 6 sanaa. Luokkaan kuuluvat lyhyet *UO*- tai *ie*-loppuiset sanat, joiden monikkomuodoissa diftongin ensikomponentti katoaa tunnuksen iin (*soihin*) edeltä. Myös nämä sanat ottavat yksikön illatiivin allomorfin *hVn*.

Luokat 20 ja 21 (*filee, rosé*) sisältävät niinikään vartalon pitkän vokaaliaineksen lyhenemät ja ovat samoja sillä erotuksella. Näistä 20 sisältää ensi sijassa mukautuslainoja, joista pitkä vokaaliaines käy ilmi, 21 sitaattilainoja, josta pitkä vokaaliaines ei käy ilmi. Allomorfeilta nämä eroavat sillä, että 20-luokkaan kuuluvat

myös illatiivit *seen* (*fileeseen*) ja *siin* (*fileisiin*) sekä monikon genetiivi *iden* (*fileiden*), mutta luokkaan 21 vain *hVn* (*roséhen*), *ihin* (*roséihin*) ja *itten* (*roséitten*).

Luokkaan 22 (*parfait*) on sijoitettu toteutuksen kannalta hankala joukko vierassanoja, joiden kirjoitusasu päättyy konsonanttiin ja ääntöasu vokaaliin, ja jotka taivutetaan puolilainausmerkin ja ääntöasun mukaisen päätteen kanssa (*parfait'na : parfait'ina*). Sanojen käsittelyn hankaluus seuraa siitä, että kirjoitusasu ei sinänsä suoraan paljasta ääntöasua, eikä Nykysuomen sanalistassa olevista tiedoista sitä selviä. Tässä luokassa taivutuksen allomorfit ovat kuten luokassa 21:kin, joten yksikön illatiivin vokaalia (*parfait'hen*) ei voida tietää, eikä tietenkään myöskään vokaalisointua.

Luokat 23—26 sisältävät konsonanttivartaloon liittyvät yksikön partitiivin ja monikon genetiivin allomorfit *tA* ja *ten*. Lisäksi luokassa 25 esiintyy nasaaliassimilaatiosta johtuva *m : n(t)* vaihtelu näiden muotojen kanssa. Vartalovokaalivaihtelun osalta kyse on luokan 7 *i : e* -vaihtelusta. Luokkien 24 ja 26 välinen ero on ainoastaan niiden monikon genetiivin allomorfien keskinäinen järjestys.

Luokkaan 24 kuuluu myös sana *meri* ja luokkaan *26* sana veri, jotka taipuvat muutoin odotuksenmukaisella tavalla, mutta niiden yksikön partitiivin tunnuksessa käytetään odotuksenvastaista sointuvokaalia (*merta* ja *verta* vastaavasti).

Luokat 27—31 sisältävät ti-si-muunnokseen liittyvät monimutkaiset vaihtelut (ks. tarkemmin esim. Karlsson (1982)), joihin sisältyy myös vartalovokaalin *i : e* -vaihtelu sekä konsonanttiklusterien yksinkertaistumat.

Luokan 27 (*käsi*) yksikön nominatiivin vartalo päättyy *si*, muiden yksikkömuotojen vartalo on astevaihtelullinen *te : de* (*kätenä : kädessä*), yksikön genetiivi *ttA* (*kättä*) ja monikkomuotojen *s*, josta vokaali on kadonnut monikon tunnuksen *i*:n edeltä kuten luokissa 5 ja 6. Luokassa 27 ovat myös *viisi* ja *kuusi*, jotka ovat numeraaleja.

Luokassa 28 (*kynsi*) *si*-sarjaa edeltää *l*, *r* tai *n*, joka aiheuttaa astevaihdellun *de*-vartalotyypin assimiloitumisen *le*-, *re*- ja *ne*-tyyppisiksi (*kynnessä*) vastaavasti.

Luokat 29 ja 30 (*lapsi ja veitsi*) ovat muutoin kuin luokat 27 ja 28, mutta *psi*- ja *tsi*-päätteet muuttuvat konsonanttivartaloisessa yksikön partitiivissa ja monikon genetiivissä *stA*- (*lasta, veistä*) ja *sten*-muotoihin (*lasten, veisten*), vastaavasti, historiallisen konsonanttiklusterin yksinkertaistuman seurauksena (Remes, 2004).

Luokka 31 (*kaksi*) sisältää *äännevaihteluina ti-si*-muunnoksen lisäksi *k : h* -vaihtelun (*kahden*), joka on seurausta vanhasta ks : ht -muutoksesta (vrt. lahti < vanh. laksi, « ruots. lax) (Remes, 2004). Muutoin luokka on kuin 27. Luokka sisältää vain sanat *yksi* ja *kaksi*, joka kattaa sen *äännerakennekuvauksen* sinänsä.

Luokka 32 (*sisar*) sisältää *r*-, *l*- ja *n*-loppuiset konsonanttivartalot, jotka vaih-

televat *e*-loppuisen vokaalivartalon (eli *re*, *le* ja *ne* vastaavasti) kanssa yksikkömuodoissa. Taivutustunnusten allomorfeiksi kuuluvat konsonanttivartalojen yksikön partitiivi *tA* (*sisarta*) sekä monikon genetiivit *ten* (*sisarten*) ja *ien* (*sisarien*). Myös numeraali *kymmenen* on sijoitettu tähän luokkaan, kuitenkin siten, että oletettu taivutusvartalo olisi *kymmen*.

Luokka 33 (*kytkin*) sisältää *n*-loppuiset konsonanttivartalot, jotka vaihtelevat vokaalivartalossa m:n kanssa (*kytkimen*). Taivutustunnusten allomorfit ovat kuten luokassa 32.

Luokka 34 (*onneton*) koostuu *tOn*-loppuisista adjektiivin abessiivin karitiivijohdosten kiteytyneistä muodoista, jotka ovat taas sanakirjan kannalta juurestaan niin etäisiä että ne on merkitty omiksi muodoikseen. Muutoin taivutus kuuluu luokan 35 kanssa (*lämmin*) *n : mA* -vaihteluun. Luokat 34 ja 35 eroavat allomorfeiltaan vain monikon genetiivin harvakäyttöisten varianttien osalta siten, että 34 voi ottaa päätteen *ten* (*onnetonten*) ja 35 *in* (*lämpimäin*).

Luokka 36 (*sisin*) sisältää kiteytyneet adjektiivien superlatiivimuodot, ja taipuu samoin kuin adjektiivin superlatiivi normaalisti *n : mPA*-vaihtelullisena. Myös luokka 37 (*vasen*) taipuu samoin ja käytännössä samoilla allomorfeilla, eroksi jää kantasana, ja teoriassa harvinaisen yksikön partitiivin allomorfin esiintyminen, sillä sanakirjan luokiutuksen mukaan superlatiivin kiteytymiin ei saa liittää pääteallomorfia *(mp)AA* (*?sisimpää, vasempaa*).

Luokka 38 (*nainen*) sisältää *nen*-loppuiset, alkujaan johtimelliset sanat, joiden taivutus perustuu johtimettomaan muotoon, eli vartalona on *se*, jonka vartalopääte se toimii kuten luokassa 29—31. Merkillepantavaa tässä luokassa on, että nominatiivin sijasta yhdyssanan määriteosan muotona käytetään nominatiivitapauksissa aina erillistä *s*-loppuista yhdyssanamuotoa (*nais-*). Muissa luokissa erillinen yhdyssanamuoto on harvinainen.

Luokat 39—42 sisältävät *s*-loppuisten äännevaihtelut. Luokan 39 (*vastaus*) *s* vaihtelee vokaalivartaloon *kse*-muodossa, jonka *e* käyttäytyy kuten vartalovokaali *e* luokassa 5. Lisäksi konsonanttivartalot saavat yksikön partitiivin allomorfinsa *tA* ja monikon genetiivin *ten*. Luokka 40 (*kalleus*) käyttäytyy monikon osalta kuin 39, mutta yksikön vokaalivartalo kuten *te : de* luokassa 27.

Luokka 41 (*vieras*) sisältää vaihtelun *s : V*, missä *s* siis esiintyy yksikön vokaalivartalossa edeltävän vokaalin pidentymänä (*vieraana*). Lisäksi monikkomuodon allomorfit ovat vokaalivartalopohjaisesti genetiivin *iden-itten* ja partitiivin *itA* sekä illatiivin *isiin*. Luokka 42 (*mies*) sisältää poikkeuksellisen *s : he* vaihtelun, muutoin vaihtelu on kuten luokissa 32 ja 33.

Luokat 43—44 ovat *t*-loppuisten sanojen vaihteluluokat, näistä 43 (*ohut*) vaihtelee *t : (∅)e*, mutta muutoin kuten luokat 32, 33 ja 42. Monikon allomorfeina vain

vokaalivartaloiset *iden*, *itten*, *itA* ja *isiin*. Luokka 44 (*kevät*) vaihtelee vokaalivartalossa vokaalin pitenemän kanssa kuten luokka 41:kin, ja näiden välisenä erona on vain loppukonsonantti.

Luokka 45 (*kahdeksas*) sisältää järjestysluvut, joiden taivutus noudattaa epäsäännöllistä *s : nte* -vaihtelua, ja jonka *t* monikkomuodoissa vaihtelee *s*:n kanssa, koska joutuu *i*:n edelle vartalovokaalin *e*:n kadon myötä.

Luokka 46 (*tuhat*) sisältää lukusanan *tuhat*, jonka taivutus on noudattaa epäsäännöllistä *t : nte* -vaihtelua, joka luokan 45 sisältää *ti : si* -vaihtelun monikossa.

Luokka 47 (*kuollut*) koostuu nut-partisiippien aktiivien muodoista, jotka ovat tarpeeksi kiteytyneitä kuuluakseen sanakirjaan. Vartalon vaihtelut koostuvat siis *ut : eet : e* -muodoista, joita edeltää nut-partisiipin tunnuksen assimiloituva *n*.

Luokka 48 (*hame*) sisältää *e*-loppuiset sanat, joihin on kielihistoriallisesti kuulunut loppukonsonantti, ja jotka taipuvat kuin niissä olisi loppukonsonantti paikalla. Myös sanan ääntöasussa on loppukonsonanttia merkitsemässä vielä jäännöslopukepiirre. Yksikön vokaalivartalossa jäännöslopuke vaihtelee pitkän vokaalin kanssa kuten luokissa 44 ja 41. Sanan yksikön partitiivin allomorfi on *ttA* kuin konsonanttivartalossa, mutta monikkoallomorfit ovat vokaalivartaloiset *iden*, *itten*, *itA*, *isiin* ja *ihin*. Luokkaan kuuluu myös uusia sanoja, jotka on muodostettu e^x-johtimella, vaikka niihin ei muutoin jäännöslopukepiirrettä kielihistorian kautta tulisikaan.

Luokkaan 49 (*askel* $\sim$ *askele*) kuuluvat sanat, joilla on kaksi erillistä taivutusvartaloa, toinen luokkaan 48 kuuluvan kaltainen *le*-, *re*- tai *ne*-loppuinen, toinen luokkaan 32 kuuluvan kaltainen *l*-, *r* tai *n*-loppuinen. Nykysuomen sanalistassa sanoista on yleensä molemmat versiot, mutta mahdollinen astevaihtelu on kuvattu vain konsonanttivartaloisille nominatiiveille (esim. *manner'*$^{49-J}$ $\sim$ *mantere*^{49}).

Yhdysnominien luokat 50 ja 51 sisältävät kaikki yhdyssanat, joita ei tämän tutkielman aiheena käsitellä tarkemmin, mutta käytännössä on niin, että luokan 50 sanoja taivutetaan kuten yhdyssanan jälkiosaa sen omassa luokassaan taivutettaisiin, ja luokan 51 sanoja taivutetaan kuten sen molempia osia omissa luokissaan. Käytännössä näitä luokkia ei suoraan itse yhdyssanojen yhteydessä sanalistasta yleensä löydy, eikä yhdyssanarajan jakopaikkaa —, joka on kieliteknologiselle järjestelmälle osin ongelmallista. Varsinaisesti sanalistassa luokan 50 sanoja, joiden osat sanalistasta löytyvät itsenäisinä tietueina ei ole luokiteltu mihinkään luokkaan, ja ne luokan 50 yhdyssanat, joiden perusosa puuttuu sanalistasta itsenäisenä tietueena, on pyritty sijoittamaan siihen luokkaan, jonka mukaan varsinainen taivutus on. Luokkaan 51 kuuluu perinteisesti adjektiivialkuisia yhdyssanoja, ja adjektiivialkuisista sanoista kaikki onkin luokiteltu joko luokkaan 50 tai 51, poikkeuksena nominialkuisista. (Eronen, 1994, sanakirjan osalta)

Luokka 52 (*sanoa*) on verbien yksinkertaisin ja vaihteluttomin luokka vastaavasti kuin luokka 1 on nomineiden kohdalla. Se sisältää vartalovokaalit O ja U, jotka siis verbien tapauksessa sanakirjamuodossa esiintyvät ennen a-infinitiivin tunnusta, tässä *A*:ta.

Luokat 53—55 sisältävät vartalon *A*:n kadon menneen ajan tunnuksen *i*:n edeltä, ja kuvaavat siitä syntyvän *ti*-sarjan *ti*-*si*-äännemuunnoksen eri vaiheita.

Luokassa 53 (*muistaa*) muunnosta ei tule, vaan menneen ajan muodossa on aina *t* (*muisti*). **Luokka 54** (*huutaa*) sisältää myös pakollisen muunnoksen menneen ajan muodossa (*huusi*). **Luokassa 55** (*soutaa*) *ti*- ja *si*-muodot ovat vapaassa vaihtelussa (*souti* $\sim$ *sousi*).

Luokka 56 (*kaivaa*) sisältää vartalovokaalivaihtelun *A : O* menneen ajan muodoissa (*kaivoi*). **Luokassa 57** (*saartaa*) luokan 56 *A : O* -vaihtelullinen sekä luokan 54 *ti : si* -vaihtelullinen *A :* $\emptyset$ -muoto ovat vapaassa vaihtelussa menneen ajan muodoissa (*saarsi* $\sim$ *saartoi*).

Luokat 58—60 ovat *e :* $\emptyset$ -vaihtelun sisältäviä luokkia, näistä **luokka 58** (*laskea*) on muutoin vaihteluton. **Luokka 59** (*tuntea*) sisältää *ti : si* -vaihtelun menneen ajan muodossa (*tunsi*) — sekä *t*:n heikon asteen assimilaation *n*:ään (*tunnen*). **Luokka 60** (*lähteä*) on myös kuin 58, mutta sisältää harvinaisena *ks : ht* -vaihtelun menneen ajan muodoissa (*lähti* $\sim$ *läksi*) kuten nominiluokassa 31.

Luokka 61 (*sallia*) sisältää vartalovokaalin *i* kadon menneen ajan (*sallin*) sekä konditionaalin (*sallisin*) tunnuksen *i*:n edeltä.

Luokissa 62—65 on pitkä vokaalivartalo, jonka jäljessä esiintyy a-infinitiivin tunnus *dA*. Vartalon pitkä vokaaliaines esiintyy lyhyenä menneen ajan ja konditionaalin tunnuksen *i*:n edellä.

Luokat 62 (*voida*) ja **63** (*saada*) eroavat nähtävästi vain sillä, että vartalon vokaaliaines on *i*-loppuinen diftongi ja pitkä vokaali vastaavasti. **Luokka 64** (*juoda*) sisältää vartalovokaalinaan väljenevän diftongin, joka käyttäytyy kuten nominiluokassa 19, diftongin ensikomponentti katoaa taivutustunnuksen *i*:n vaikutuksesta. **Luokka 65** (*käydä*) sisältää poikkeuksellisen *y*:n vaihtelun *v*:ksi menneen ajan (*kävin*) tai konditionaalin (*kävisin*) tunnuksen *i*:n edeltä.

Luokka 66 (*rohkaista*) sisältää *s*-konsonanttivartalot ja **luokka 67** (*tulla*) *l*- ja *r*-konsonanttivartalot. Konsonanttivartaloihin liittyy potentiaali ja nut-partisiippi tunnuksen *n* assimiloituneena. Luokkaan 67 on pantu myös *olla*-verbi, johon sisältyy sekä poikkeavaa taivutusta kolmansien persoonien muodoissa että suppletiivinen paradigma potentiaalimuodoissa.

Luokkaan 68 (*tupakoi* $\sim$ *tupakoitsee*) kuuluvat luokan 62 sanat, joilla on harvinaisena vaihtoehtoisena taivutusvartalona tyypin 69 vartalo.

Luokka 69 ja 70 (*valita, juosta*) sisältävät *e : ∅* -vaihtelun sekä kompleksisemman konsonanttiklusterin yksinkertaistuman siten, luokkaan 69 kuuluu *tse : ts* -vartalot ja 70 *kse : st : s* -vartalot.

Luokkaan 71 (*nähdä*) kuuluu *ke : h*-vaihtelu.

Luokka 72 (*vanheta*) sisältää vartalot, joissa *ne*-osa on ilmeisesti perua kokonaan toisesta paradigmasta, eli vartalon vaihtelu on *∅ : ne*.

Luokka 73—75 sisältävät tavallisen vokaalivartalon ja lisäksi menneen ajan tunnusta edeltää vaihtelun johdosta *s*. Luokat eroavat toisistaan siten, että *a*-vartaloinen **73** (*salata*) ottaa menneen ajan tai konditionaalin tunnuksen vokaalivartalon lyhyen vokaalin jälkeen (*salasin*). *E*-, *O*- tai *U*-vartaloinen **74** (*katketa*) säilyttää pitkän vokaaliaineksen ennen konditionaalin tunnusta (*katkeaisin*), mutta konditionaalissa tapauksittain myös luokan 73 mukainen kato on mahdollinen (*katkeisin*). *I*-vartaloinen **75** (*selvitä*) taipuu kuten 74, mutta luokan 73 mukaista konditionaalia ei esiinny, ja se menisikin yhteen konditionaalimuodon kanssa.

Luokka 76 (*taitaa*) on 58, mutta lisäksi se sallii n-konsonanttivartaloiset muodot potentiaalissa (*tainnee*) ja nut-partisiipissa (*tainnut*).

Luokkaan 77 (*väräjää*) on koottu *AjAA*-tyypin verbit, joista esiintyy käytössä vain yksikön kolmannen indikatiiveja (*väräjää : väräji*) ja konditionaaleja (*väräjäisi*).

Luokassa 78 (*kaikaa*) on verbit, joista esiintyy vain yksikön kolmannen indikatiivin menemättömän ajan muoto (*kaikaa*) ja konditionaali (*kaikaisi*).

Luokkiin 99 ja 101 on pantu adverbejä, partikkeleja ja pronominejä, joita en tässä tutkielmassa käsittele.

Astevaihteluluokkia aineistossa on 13, jotka kuvaavat astevaihtelun alaisena olevan äänneparin, mutteivät astevaihtelun suuntaa, joka käy ilmi taivutusluokasta sinänsä. Astevaihtelut on numeroiden sijasta annettu kirjaimin A—M, ja ne muodostavat kombinaatioita numeroluokkien kanssa. Luokat on järjestetty siten, että kirjaimet A—C kuvaavat kvantitatiivisen astevaihtelun eri klusiileille, ja loput D—M kuvaavat kvalitatiivisen astevaihtelun eri toteutumavariantteja.

Astevaihteluluokista lienee huomionarvoista, että uudehkolla vierasperäisten konsonanttien *g*, *b* ja *d* kvantitatiivisella astevaihtelulla (*dubata : dubbaan* ja *digata : diggaan*) ei ole omaa astevaihteluluokkaa, eikä sitä ole Nykysuomen sanalistaan merkitty ja painetussa sanakirjassa se löytyy vain sana-artikkelin proosasta (Eronen, 1994). Myös sanassa *auer* esiintyvä *t : ∅* -astevaihtelu puuttuu luokituksesta.

Astevaihtelun kannalta on merkittävää, että astevaihtelun toteutuminen sanakirjamuodossa riippuu siitä, onko astevaihtelun alainen tavu sanakirjamuodossa lyhyt avotavu vai ei. Perinteisesti kieliopissa sanakirjamuodoltaan avotavuisia eli

vahva-asteisia muotoja on kutsuttu astevaihtelun pääsäännön suoran astevaihtelun alaisiksi, ja niitä, joissa on umpitavu ja heikko aste sanakirjamuodossa käänteisen astevaihtelun mukaisiksi (Hakulinen et al., 2004). Nomineilla suoraan astevaihteluun kuuluvat yksinkertaisesti taivutusluokat 1—31, jotka ovat vokaaliloppuisia nomineja. Luokat 32—47 ovat konsonanttiloppuisia käänteisen astevaihtelun sisältäviä luokkia. Verbien kohdalla suoraan astevaihteluun kuuluvat 52—65, käänteiseen 66—75. Kun astevaihtelun alainen äänne selviää sanaluokkanumeron ja astevaihtelukirjaimen avulla, sen paikantaminen selviää lähes aina triviaalisti sanakirjamuodon kirjoitusasusta poimimalla viimeinen eli oikeanpuoleisin kyseeseen tuleva kirjain. Poikkeuksia luovat muutamat verbiluokat, joissa a-infinitiivin tunnuksen *t* pitää sivuuttaa ensin sekä käänteisessä astevaihtelussa luokan D selvittäminen, kun siinä sanakirjamuodossa *k*:ta vastaa kato.

Taivutusluokituksen jäsentimessä käyttämistäni piirteistä olen laatinut kuvaukset taulukkoihin 1 nomineille ja 2 verbeille, sekä astevaihtelun osalta yhteenvetona taulukkoon 3. Samat tiedot löytyvät teoriassa Nykysuomen sanalistan (Kotimaisten Kielten Tutkimuskeskus, 2006) mallitaivutustaulukoista hieman eri muodossa.

Nominitaulukon äännevaihtelusarakkeessa kuvataan vartalossa esiintyvät äännemuutokset poislukien astevaihtelusta johtuvat —, jotka kuitenkin saattavat toisinaan olla luokalle ominaisia. Sarakkeessa äännerakenne on esitetty yleistäen se äännerakenne, jota kaikki luokan sanat (yksikön nominatiivissa) noudattavat, sanan lopusta lukien, eli äännerakennekuvaus ”u” tarkoittaisi vartalovokaalia *u* ja ”psi” vain *psi*-loppuisia vartaloita. Monissa luokissa on muutamia sanoja, jotka poikkeavat odotuksenmukaisesta muodosta, esimerkiksi osittain mukautuneita vierassanoja. Tällaiset tapaukset on usein merkitty sulkeilla. Esimerkiksi luokkaan 3 näyttävät kuuluvan ensi sijassa *io*- ja *iö*-loppuiset omaperäiset sanat, mutta myös muut vokaaliyhtymään päättyvät uudet vierasperäiset sanat (esim. *zombie*[3], sikäli kun se lausutaan suunnilleen [ˈtsom.bi.e]).

Huomattavaa on myös, että jotkin pronominit ja numeraalit on luokiteltu mukaan luokkiin siten, etteivät ne sinänsä täsmää kuvaukseeni, vaikka etymologisesti sopivatkin, esimerkiksi sanan *kumpi* kieliopillistuneet kliitilliset muodot, kuten sanaksi merkitty *kumpikaan*, ovat luokassa 16 muiden komparatiivien kanssa, vaikka ne taipuvat kliittinsä edestä (*kumpikaan : kummatkaan : kummillekaan*).

Taivutusmuotosarakkeissa yks. ptv., yks. ill., mon. ptv., mon. gen. ja mon. ill. on kuvattu monta eri allomorfia sisältävien taivutuspäätteiden vapaassa vaihtelussa olevat vaihtoehdot kuten ne on lähdeaineistossa annettu. Tästä lienee huomattava, että aineiston mukaan annetulla järjestykselläkin on väliä, ja tosiaan esimerkiksi luokkien 24 ja 26 ainoa morfologinen ero on kahden taivutustunnuksen järjestyksen ero.

Allomorfien ja vartalotyyppien suhdetta en ole kuvannut, sen yksityiskohdat löy-

tyvät vain toteutuksesta tarkemmin. Vartalotyyppeihin viittaavat nimet nominatiivivartalo, heikko vokaalivartalo, vahva vokaalivartalo, konsonanttivartalo ja monikkovartalo ovat kuvauksissa kuten perinteisissä koulukieliopeissa, eli yksikön nominatiivi sellaisenaan, yksikön essiivin vokaalivartalo, monikon nominatiivin vokaalivartalo, yksikön partitiivin konsonanttivartalo ja *i*-tunnuksisten monikkomuotojen *i*:tä edeltävä vokaalivartalo vastaavasti.

Taulukon arvaukset perustuvat morfologian kehityksessä käytettyihin arvioihin, joiden lähteenä ovat olleet aiemmin mainitut kieliopit (Remes, 2004; Setälä, 1930), Nykysuomen ja Perussanakirjan kuvaukset (Eronen, 1994, 1997), aiemmin toteutetut kieliteknologiset morfologiajärjestelmät (Koskenniemi, 1983) sekä Nykysuomen sanalistan silmämääräinen ja koneellinen tarkastelu. Eräänä työkaluna vartalon äännerakennevaihteluita arvaillessani olen käyttänyt työkalua, joka hakee vapain GNU GPL -sovelluksin luettelosta tietyn luokan sanoja, tietyn vartaloisia sanoja, sekä tietyn luokan sanoja, jotka eivät ole kirjoitusasultaan tiettyä hakulauseketta vastaavia. Työkalua käytetään hakemalla kaikki luokan sanat listasta, arvaamalla äännerakenne, ja pyytämällä kaikkia arvatusta äännerakenteesta poikkeavia luokan sanoja, ja jos tuloksena on tyhjä joukko on ainakin selvää että hakulauseke kattaa koko listassa olevien luokan sanojen äännerakenteen. Tästä seuraa, että monet tutkielmassa käytetyistä arvioista sanojen luokkien äännerakenteesta ovat vailla varsinaista lingvististä perustetta, eli silmämääräisiä. Osa arvauksista on toki myös pyritty liittämään em. kielioppien alkeiskuvauksiin kielen äännekehityksestä, kuten on aiemmissa luokkakuvauksissa huomattavissa, mutta systemaattisesti tätä ei ole tehty.

Taulukko 1: Nominiluokkien äänne- ja muotorakenne sekä muut erottavat piirteet

Piirre→ ↓Luokka	**Vartalo-vaihtelu**	**Äänne-rakenne**	**Yks. ptv.**	**Yks. ill.**	**Mon. ptv.**	**Mon. gen.**	**Mon. ill.**
Vokaalivartaloiset, suoralla astevaihtelulla							
1	—	O tai U	A	Vn	jA	jen	ihin
2	—	O tai U $\geq$ 3 tavua	A	Vn	jA, itA	jen, iden, itten	ihin
3	—	iO (oe, yo ie, eo, ...)	tA	Vn	itA	iden, itten	ihin
4	—	kkO $\geq$ 3 tavua	A	Vn	jA, itA	jen, iden, itten	ihin, ihin‡
5	i† : e : $\emptyset$	i	A	Vn	jA	ien	ihin
6	i† : e : $\emptyset$	i	A	Vn	jA,	ien,	ihin

(jatkuu seuraavalla sivulla)

Piirre→ ↓Luokka	**Vartalo-vaihtelu**	**Äänne-rakenne**	**Yks. ptv.**	**Yks. ill.**	**Mon. ptv.**	**Mon. gen.**	**Mon. ill.**
					itA	iden, itten	
7	i : e : ∅	i	A	Vn	iA	ien	iin
8	e	e	A	Vn	jA	jen (in)	ihin
9	A : O	A	A	Vn	jA	jen (in)	ihin
10	A : ∅	A (An)	A	Vn	iA	ien (in)	iin
11	A : O : ∅	A, ≥ 3 tavua	A	Vn	iA, itA (jA)	ien, iden, itten (jen, in)	iin, ihin
12	A : O	A, ≥ 3 tavua	A	Vn	itA	iden, itten (in)	ihin
13	A : O	A	A	Vn	itA, jA	jen, itten, iden (in)	ihin
14	A : O	kkA, (ttA, ppa) ≥ 3 tavua	A	Vn	itA, jA	iden, itten, jen	ihin, ihin‡
15	A : ∅	eA (UA, OA)	A, tA	Vn	itA	iden, itten (in)	isiin ihin
16	mpi : mPA	mpi	A	Vn	iA	ien (in)	iin
17	$V_1V_1 : V_1$	pitkä vokaali	tA	seen	itA	iden, itten	isiin (ihin)
18	VV : V	pitkä vo-kaaliaines	tA	hVn	itA	iden, itten	ihin
19	$V_1V_2 : V_2$	laveneva diftongi	tA	hVn	itA	iden, itten	ihin
20	$V_1V_1 : V_1$	pitkä vokaali	tA	hen, seen	itA	iden, itten	ihin, isiin
21	vierasperäiset pitkään vokaaliainekseen loppuvat						
22	vierasperäiset kirjoituksessa konsonanttiin, ääntäessä vokaaliin loppuvat						
23	i : e	i	tA	Vn	iA	ien	iin
24	i : e	i	tA	Vn	iA	ien, ten[1]	iin
25	mi : me : n	mi	tA, A	Vn	iA	ien, ten	iin
26	i : e	i	tA	Vn	iA	ten, ien[1]	iin
27	si : te : t	si	tA	Vn	iA	ien	iin

(jatkuu seuraavalla sivulla)

Piirre→ ↓Luokka	**Vartalo-vaihtelu**	**Äänne-rakenne**	**Yks. ptv.**	**Yks. ill.**	**Mon. ptv.**	**Mon. gen.**	**Mon. ill.**
						(ten)	
28	si : te : t	nsi, Lsi	tA	Vn	iA	ien (ten)	iin
29	psi : pse : s (ksi : kse : s)	psi (ksi)	tA	Vn	iA	ten, ien	iin
30	tsi : tse : s	tsi	tA	Vn	iA	ien (ten)	iin
31	ksi : hte : h	yksi, kaksi	tA	Vn	iA	ien	iin
konsonanttivartalot, käänteisastevaihtelulla							
32	L : Le n : ne	L n	tA	Vn	iA	ien, ten	iin
33	n : me	in (n)	tA	Vn	iA	ien, ten	iin
34	n : mA	tOn	tA	Vn	iA	ien, (ten)	iin
35	mmin : mpimä	lämmin	tA	Vn	iA	ien (in)	iin
36	n : mpA	in,	tA	Vn	iA	ien, ten (in)	iin
37	n : mpA	vasen	tA (A)	Vn	iA	ien, ten (in)	iin
38	nen : se	nen	tA	Vn	iA	ten, ien	iin
39	s : kse	s	tA	Vn	iA	ten, ien	iin
40	s : te : *t : kse	s	tA	Vn	iA	ien	iin
41	s : V	s	tA	seen	itA	iden, itten	isiin (ihin)
42	s : he	mies	tA	Vn	iA	ien	iin
43	t : e	Ut	tA	Vn	itA	iden, itten	isiin, ihin
44	t : V	ät	tA	seen	itA	iden itten	isiin (ihin)
45	s : nte : t : ns	s	tA	Vn	iA	ien	iin

(jatkuu seuraavalla sivulla)

Piirre→ ↓Luokka	**Vartalo-vaihtelu**	**Äänne-rakenne**	**Yks. ptv.**	**Yks. ill.**	**Mon. ptv.**	**Mon. gen.**	**Mon. ill.**
46	t : nte : ns	tuhat	tA	Vn	iA	ien (ten)	iin
47	Nut : Nee	nUt, LUt, sUt	tA	seen	itA	iden, itten	isiin, ihin
48	e : ee	e^x (ori, kiiru)	ttA	seen	itA	iden, itten	isiin, ihin
49	l ∼ le, r ∼ re, n ∼ ne	l ∼ le r ∼ re n ∼ ne	tA	(Vn) seen	iA, itA	ien, ten iden, itten	iin, ihin, isiin

† myös konsonanttiloppuiset vierassanat, joissa sidevokaali i, mm. kaikki konsonanttiloppuiset sitaattilainat

‡ ihin käy sekä vahvaan että heikkoon vartaloon

[1] luokat 24 ja 26 eroavat toisistaan vain näiden allomorfien *järjestyksen* perusteella

Taulukko verbien vaihtelulle on hieman yksinkertaisemman näköinen, sillä verbeissä ei esiinny taivutuspäätteiden allomorfien vapaata vaihtelua siten kuin nomineissa. Sen sijaan verbeillä esiintyy vain vartalon vaihtelua, joka liittyy joko vain menneen ajan muodon i-tunnukseen tai sekä siihen että konditionaaliin, riippuen kielihistoriallisista seikoista, joihin ei tässä tarkemmin puututa. Lisäksi merkillepantavaa on sanakirjamuodon suhde varsinaiseen verbivartaloon, sillä sanakirjamuodon eli a-infinitiivin tunnuksena esiintyy variantit *A*, *tA*, *dA* ja *(tA)A*, jotka eivät kuulu varsinaiseen taivutusvartaloon. A-infinitiivin allomorfi selviää kuitenkin sarakkeesta äännerakenne, sillä se kuvaa nimenomaan sanakirjamuodon äännerakennetta. Sarakkeessa pret. eli preteriti on menneen ajan muodosta kuvattu sen vaikutus vartaloon, eli esim. mahdollinen ti>si-muunnos t-loppuisissa vartaloissa. A-infinitiivin allomorfin mukaan määräytyvät infinitiivien tunnuksessa mahdollisesti esiintyvä *d*, joka on kuvattu sarakkeessa inf. Kyseessä on sinänsä vain tavan astevaihtelukuvio, sillä poikkeuksella, että heikon asteen tunnuksesta *ð* on tietyissä konteksteissa jäänyt jäljelle nykysuomessa kato (ks. tarkemmin Hakulinen (1979)). Vastaavasti *s*:n jälkeisessä asemassa aste on vahva. Sarakkeessa pass. on kuvattu se vaihtelu, joka määrittää onko t-alkuisen passiivin alussa yksi *t* (nähty, syötiin) vai kaksi *t*:tä (tapettu, haudattiin). Nämä vaihtelut ovat käytännössä samat kuin väitöskirjassa Koskenniemi (1983) kuvatun järjestelmän D- ja Z-morfofoneemit.

Taulukko 2: Verbiluokkien äänne- ja muotorakenne sekä muut erottavat piirteet

Piirre→ ↓Luokka	**Vartalo-vaihtelu**	**Äänne-rakenne**	**Pret.**	**Inf.**	**Pass.**
52	—	OA, UA (iA)	i	—	t
53	A : ∅	tAA (AA)	i	—	t
54	tA : s∅(i)	tAA (sAA)	si	—	t
55	tA : t∅(i) ∼ si	tAA	si	—	t
56	a : o	aa	i	—	t
57	taa : s∅(i) ∼ : o	Caa(r)taa	si	—	t
58	e : ∅	eA	i	—	t
59	nte : ns∅(i)	tuntea	si	—	t
60	hte : ks∅(i)	lähteä	si	—	t
61	i : ∅	iA, yä	i	—	t
62	i : ∅	idA	i	d	t
63	$V_1V_1 : V_1$	pitkä vokaali+dA, 1 tavu	i	d	t
64	$V_1V_2 : V_2$i	UOdA, iedA, 1 tavu	i	d	t
65	y : v	käydä	i	d	t
66	s : se	stA	i	t	t
67	L : Le n : ne	LLA nnA	i	l	t
68	i : ∅ ∼ ts	OidA ∼ OitseA	i	d	t
69	t : tse	itA	si	t	tt
70	s : kse	UOstA, iestA	si	t	t
71	h : ke	hdA	i	d	t
72	t : ne	tA	i	t	tt
73	t : s	AtA	si	t	tt
74	t : s	OtA, UtA, etA	si	t	tt
75	t : s	itA	si	t	tt
76	t : s	taa	si	—	tt
77	a : ∅	AjAA, ≥ 3 tavua	i	—	—
78	—	AA	—	—	—

Astevaihtelun osalta taulukko on yksinkertainen, ja löytyy liki sellaisenaankin aineistosta Kotimaisten Kielten Tutkimuskeskus (2006). Astevaihtelukirjain sinänsä on morfofonologian kannalta redundanttia, sillä kun tiedetään vaihtelu ja taivutusluokasta vaihtelun suunta, kyseeseen tuleva äänne selviää lopputavun alusta ja sen toisen asteen vastine kontekstista säännönmukaisesti.

Taulukko 3: Astevaihteluluokkien vaihteluparit

Aste→ ↓Luokka	**Vahva**	**Heikko**
A	kk	k
B	pp	p
C	tt	t
D	k	∅
E	p	v
F	t	d
G	k	g
H	p	m
I	t	l
J	t	n
K	t	r
L	k	j
M	k	v

2.3 Korpukset

Lopullisen järjestelmän vertailutesteissä käytin aineistona korpusmateriaalia, joka on saatavilla Tieteellinen laskenta CSC Oy:n ylläpitämällä palvelimella. Aineisto koostuu suomen kielen tekstipankin ei-vapaasti lisensoiduista B-korpuksista, joita sinänsä on lupa käyttää kaupallisen tai open source -sovelluksen kehittämiseen, mutta esimerkiksi aineiston siirtäminen pois CSC:n omalta palvelimelta ei tule kyseeseen, joten korpus-aineistoa ei voi esimerkiksi liittää valmiin kokonaissovelluksen osaksi esimerkiksi regressiotestausta varten. Regressiotestauksella tarkoitetaan tässä sellaista kieliteknologisen morfologiani kehitykseen liittyvää toimivuustestausta, joka kertoo toimiiko uusi versio samoin kuin aiemmat.

Sisällöllisesti näissä korpuksissa teksti on sanomalehdistä otettua, joten sen voi olettaa olevan suurelta osin laadullisesti ehjää yleiskieltä. Lisäksi korpukset on valmiiksi jäsennetty käyttäen toista, kaupallista jäsennintä, joten tulokset antavat myös karkeaa viitettä, mihin kehittämäni järjestelmä laadullisesti sijoittuu.

3 Menetelmät

Äärellistilainen morfologinen jäsennin on sellainen äärellistilainen transduktori, joka sisältää kuvaukset sanojen taivutusmuotojen ja niiden sanakirjamuotojen sekä taivutusmuotojen tunnisteiden välillä (esim. *kissasta* ↔ *kissa yksikön elatiivi*) tai päinvastoin. Kuvauksen toteuttaminen lähtee siitä, että muunnetaan XML-muotoinen sanalista sellaiseksi, jonka SFST osaa lukea transduktoriksi, ja tätä transduktoria säännöin ja äärellistilaisin operaatioin käsittelemällä saadaan aikaiseksi sellainen äärellistilainen transduktori, joka kuvaa sanakirjamuodot taivutusmuodoiksi.

Kappaleessa 2 hahmotellun perusteella äärellistilainen järjestelmä, joka kuvaa sanojen vartaloissa esiintyvät äännemuutokset, liittää sopivat taivutuspäätteet sanojen sopiviin vartaloihin ja niin edelleen on perustettavissa pelkästään sanakirjamuotoihin, taivutusluokkanumeroihin ja astevaihtelukirjaimiin, joten XML-sanalista pitää muuntaa äärellistilaisen järjestelmän sanalistaksi, jossa jokainen sana muodostuu kolmikosta sanakirjamuoto, taivutusnumero ja astevaihtelukirjain. Tämä muunnos on tehty XSLT 2.0[10] -kielellä ja toteutusta on kuvattu kappaleessa 3.2.1. Tässä vaiheessa lähdetään siis liikkeelle tietueista tyyppiä `<st><s>pata</s><hn>1</hn><t><tn>9</tn><av>F</av></t></st>` ja päädytään tyyppiin `pata<9><f>`.

Itse äärellistilainen kuvaus, joka perustuu tuotettuun sanalistaan, on joukko äärellistilaisten transduktorien operaatioita ja sääntöjä, joilla järjestelmällisesti tuotetaan sanan sanakirjamuodosta (tyyppiä `pata<9><f>`) taivutusvartalot (tyyppiä `pado<monikko>`) ja näistä taivutustunnukset liittämällä (tyypiksi `padoiss<A><pl><ine>` ja morfofonologiset (tyypiksi `padoissa`) vaihtelut säännöin toteuttamalla päädytään transduktoriin, joka sisältää halutun morfologisen kuvauksen (tyyppiä `pata<9><f><pl><ine>` ↔ `padoissa`). Tämän toteutuksen kuvauksen olen jakanut kahteen osaan: Aluksi kuvaan lyhyesti äärellistilaisten menetelmien pohjateoriaa ja esittelen käyttämiäni notaatioita kappaleessa 3.1. Kappaleessa kuvatuin merkinnöin ja käytännöin itse järjestelmä on kuvattu kappaleessa 3.2.

Jäsentimen kattavuuden osalta on mainittava, että toteutus kattaa vain sanojen taivutuksen, eli sisältää nomineilta vain sijapäätteet, omistusliitteet ja kliitit sekä verbeiltä aika- ja tapamuodot, infinitiivit ja partisiipit, sekä persoonamuodot ja pääluokkataivutuksen. Taivutuksen oikeellisuuden perusteeksi on enimmäkseen oletettu, että Nykysuomen sanalistan luokat ovat oikein, ja kun käytetyt mallit kuvaavat osan sanoista per luokka oikein, muidenkin pitäisi kuvautua. Toteutus sisältää testaustarkoituksiin käytettävän naiivin yhdyssanamuodostussäännöstön, joka

[10] `http://www.w3.org/TR/2007/REC-xslt20-20070123/`

muodostaa mielivaltaisen paljon perusyhdyssanoja, eli nomineista muodostuvia yhdyssanoja, joiden määriteosa on joko yksikön nominatiivissa tai genetiivissä ja vain jälkiosa taipuu.

Testitarkoituksiin valitut ei-avoimet, epävapaat korpukset ovat myös XML-aineistoa, joten niitäkin on käsitelty XSLT-pohjaisin menetelmin, lisäksi niistä XSLT-menetelmillä irroitettua dataa on verrattu SFST:n morfologisiin jäsennyksiin Python-pohjaisilla menetelmillä, josta lyhyt kuvaus kappaleessa 3.3.

3.1 Äärellistilaisten menetelmien teoriasta

Tässä kappaleessa kuvataan lyhyesti äärellistilaisten menetelmien teoriaa siltä osin kuin se on tarpeellista suomen morfologian toteutuksessa ja samalla esitellään SFST-toteutuksen tapaa merkitä äärellistilaisia operaattoreita ja ilmauksia. Laajempia ja kokonaisempia kuvauksia aiheesta löytyy vaikka teoksista Aho et al. (2007); Beesley ja Karttunen (2004).

Äärellistilainen transduktori on teoreettinen malli, joka kuvaa äärellistilaisten kielten relaatiota kuusikolla $(\Sigma, \Gamma, S, s_0, F, \delta)$, jossa Σ on ylemmän kielen aakkosto, Γ alemman kielen aakkosto, S on äärellinen joukko tiloja, s_0 alkutila joukosta S, F S:n osajoukko lopputiloja ja $\delta : S \times (\Sigma \cup \epsilon) \times (\Gamma \cup \epsilon) \rightarrow S$ siirtymärelaatio joka kuvaa tilalta toiselle aakkospareilla. (Jurafsky ja Martin, 2000)

Äärellistilaisten menetelmien havainnollistavassa kuvaamisessa käytetään yleensä sellaista grafiikkaa, että tiloja merkitään ympyröillä tai ellipseillä, ja tilojen välisiä siirtymärelaatioita merkitään nuolilla. Sekä tilat, että siirtymät ovat nimettyjä ja siirtymien nimet muodostuvat aakkoston olioista. Tilojen nimet kirjoitetaan tässä ympyrän sisälle, ja siirtymien nimet kaaren päälle tai sen läheisyyteen.

Säännölliset ilmaukset ovat nimitys menetelmille kuvata äärellistä automaattia kuvaamansa kielen aakkosilla, jotka ovat mielivaltaisia kielen symboleja, kuten esimerkiksi luonnollisessa kielessä konkreettisia latinalaisia aakkosia, sekä operaattoreilla, joilla mallinnetaan aakkosten tai aakkosjonojen mielivaltaisia joukko-opillisia yhdistelmiä ja toistoja. Morfologiaa käsitellessä relevantteja yksiköitä ovat tietenkin aakkoset, joilla kirjoitetaan kielen sanoja, ja lisäksi on tyypillistä käyttää joitain erikoismerkintöjä kuvaamaan sanojen taivutusmuotoja, morfofonologian abstrakteja lisämerkintöjä käsittelyn aikana ja muita erikoismerkkejä. Tyypillisimmät säännölliset ilmaukset käsittelevät yksitasoisia säännöllisiä automaatteja, tässä tutkielmassa on kyse useimmiten kaksitasoisista äärellisistä transduktoreista, joten kuvauksetkin on annettu hieman poikkeavalla SFST-PL-kielisellä murteella säännöllisistä ilmauksista. SFST-PL poikkeaa myös osin muissa transduktorijärjestelmissä käytetyistä murteista, joten kuvaan tässä tarkemmin peruso-

peraatioita ensiksi SFST-PL-kielisesti. Muihin merkintöihin tottuneet voivat löytää liitteestä B taulukon, jossa on verrattu SFST:n notaatiota yleisempiin variantteihin.

Atomisimmat peruselementit äärellisissä järjestelmissä ovat aakkoston symboleista muodostuvien jonojen joukkoja, eli formaaleja kieliä, joita tässä merkitään symboleilla tyyppiä $\$L\$$. Aakkosten yksittäiset symbolit merkitään yksinkertaisesti a ja symboliparit $a : b$.

Perusolioita äärellisissä verkoissa ovat universaalit ja tyhjät kielet (merkitään vastaavasti symboleilla $.*$ ja $<>$) sekä useimmat niiden ja mielivaltaisten kielten kombinaatio-operaattoreista, kuten ehdollisuus (so. nollasta yhteen toistoa merkitään postfiksaalisella kysymysmerkillä) ja valinnaisuus (nollasta äärettömään toistoa merkitään postfiksaalisella tähdellä).

Yhdiste (union) on loogisesti disjunktiivinen, jos siis meillä on kieli $\$L_1\$$, jossa on 'kissa' ja kieli $\$L_2\$$, jossa on 'koira', niiden yhdiste on kieli $\$L_1\$|\$L_2\$$, jossa on 'kissa' tai 'koira'. Konkatenaatio on merkkijonokäsittelyssä usein käytetty termi yksinkertaisesta peräkkäin asettamisesta, eläinkieltemme $\$L_1\$$ ja $\$L_2\$$ konkatenaatio tässä järjestyksessä olisi siis kieli $\$L_1\$\$L_2\$$, jossa on 'kissakoira'-olio. *Leikkaus* (intersection) on taas loogisesti konjunktiivinen, siis vaikkapa kielen $L_{1,2}$, jossa on 'kissa' tai 'koira' ja kielen $L_{2,3}$, jossa on 'koira' tai 'banaaniovi' leikkaus on kieli $\$L_{1,2}\$\&\$L_{2,3}\$$, jossa on vain 'koira'. *Erotus* (difference) on operaatio, jossa kielletään tai poistetaan asioita toisesta kielestä, esimerkiksi $\$L_{1,2}\$ - \$L_2\$$ olisi $\$L_1\$$.

Monipuolisemmista operaatioista äärellisille relaatioille eli transduktoreille tyypillinen *kompositio* (composition) on tässä yhteydessä kuvaus ensimmäiseltä syötteeltä toiselle tulosteelle, jos ja vain jos ensimmäinen syöte kuvautuu ensimmäiselle tulosteelle, joka kuvautuu toisena syötteenä toiselle tulosteelle mielivaltaisen (kohdistetun) polun kautta. Esimerkiksi jos meillä on relaatio, jossa on pari $a : b$ ja komposition toisena osapuolena relaatio, jossa on pari $b : c$, komposition $a : b||b : c$ tuloksena on relaatio, jossa a ja c ovat nyt pari $a : c$. Käytännön tasolla siis kompositiolla voidaan kuvata peräkkäisinä toteutumina yksittäiset säännöt siten, että välivaiheet jäävät aina pois näkyvistä. Esimerkiksi $a : b||b : c||c : d||d : e == a : e$, joka morfofonologisessa maailmassa tarkoittaa, että voidaan soveltaa järjestyksessä peräkkäin sääntöjä siten, että edellisen lopputuloksella voi olla kontekstuaalinen vaikutus seuraavan sisältöön.

Sivuutuksella (ignoring) lisätään transduktorin joka väliin mielivaltaisen monta sivuutettavaa symbolia siten, että sääntönä se voisi sivuuttaa symboleita, jos ne ovat käsittelyn kannalta irrelevantteja. Jos esimerkiksi haluaisimme tarkastella sanassa esiintyviä vokaaleja, voisimme sivuuttaa kaikki konsonantit.

Kielten fonologisia relaatioita kuvatessa säännöllisten ilmausten lisäksi käytetään sääntöjä, joista osa tunnetaan kaksitasosääntöinä (Koskenniemi, 1983). Näistä *kontekstirajoitusta* (context restriction) määrää relaatiossa tai kielessä esiintyvän symbolijonon mahdolliset esiintymäpaikat, eli jos symbolijono esiintyy, sillä on oltava operaattorin määräämä konteksti. Esimerkiksi jos kielessä on morfofonologinen vaihtelu, joka muuttaa t:n d:ksi vain, kun se on umpitavun alussa vokaalien välissä, voisimme rajoittaa parin t ja d tällaiseen ympäristöön $Vt => dVC$. Kontekstirajoitus toimii relaationa siten, että se sallii parin kuvauksen jos se esiintyy kontekstissa (kuten edellä).

Pintamuodon pakotus (surface coercion) toimii siten, että se määrää parin, jos konteksti toteutuu, eli jos vaikkapa kielessä on nasaaliassimilaatio pakollinen, voitaisiin kuvata nasaalit pareina ääntöpaikaltaan sopiviin nasaaleihin kontekstin ääntöpaikan mukaan, esim. $n <= mp$. Näiden yhdistelmä loogisesti sekä sallii parin esiintyä vain annetussa kontekstissa että pakottaa parin olemaan juuri määrätty pari kun se esiintyy annetussa kontekstissa, eli määrää parin esiintymään jos ja ainoastaan jos se esiintyy annetussa kontekstissa.

Toisinkirjoitus (replace) on operaatio, jolla kuvataan kieliä toisille ja relaatioita toisille (Karttunen, 1995). Se on esimerkiksi kaksitasosääntöjä käytännöllisempi operaatio tilanteessa, jossa relaatio, joka kuvataan, ei ole pari 1:1 vaan erimuotoinen (Schmid, 2007b), kuten esimerkiksi lavenevan diftongin ensikomponentin kato tai säilyminen vanhoissa sanoissa voitaisiin kuvata $uo : o \ \hat{} \rightarrow \ \_\_i$.

Toteuttamani järjestelmä toimii lopulta replace-sääntöjen järjestyksessä soveltamisella. Replace-sääntöjä yhdistetään sanalistasta tehtyyn transduktoriin sääntö kerrallaan komposition avulla. Replace-sääntö rakentaa aina sellaisen transduktorin, joka kompositoidessa kuvaa kaiken säännön kannalta irrelevantin identiteetikseen, ja säännön alaiset merkkiparit säännön konteksteissaan väistämättä toteutettavaksi.

Tässä sanaluokitellun listan kanssa voimme aina valita kontekstiksi taivutusluokkanumeron kerrallaan, silloin kun se on tarpeen. Tästä esimerkiksi kun tiedämme, että *vesi* ja *käsi* ovat taivutusluokan 27 sanoja, voimme kuvata vain luokalle 27 kontekstilliset toisinkirjoitussäännöt, jolla perusmuodon lopun *si* tulee yksikön vokaalivartaloissa taivutusvartaloissa olemaan *te* tai *de*. Koska kompositoidessa sääntötransduktoria oikealle ylempi taso jää paikalleen, ja alemman tason muutokset toteutuvat, säännöstön kompositoinnin jälkeen analyysitasolla on esimerkiksi *käsi* kun vartalon teon jälkeen generointitasolla voi olla *käte*.

Itse morfologian kuvauksen kannalta vähemmän relevantteja ovat äärrellistilaisten verkkojen ylläpidolliset ja käytännölliset operaatiot, kuten ϵ-poisto, determinisointi, minimointi, vertailu jne., jotka sovelluksissa ovat sinänsä tarpeellisia, ja joiden toteutuksen tehokkuus ja toiminta lopulta vaikuttavat järjestelmän suo-

rituskykyyn. Tämän tutkielman vertailussa lienee merkillepantavaa, että SFST-järjestelmässä nämä transduktorin ylläpidolliset toimenpiteet suorittaa SFST itse, eikä niitä voi sääntöjärjestelmän kirjoittaja vahvasti hallita.

3.2 Suomen kielen äärellistilaisen automaattisen morfologisen jäsentimen toteutuksesta

Tässä morfologian toteutuksessa lähdetään liikkeelle luettelosta sanakirjamuotoja taivutus- ja astevaihteluluokkineen ja päädytään transduktoriin, joka kuvaa jokaisen sanan jokaisen taivutusmuodon sanakirjamuotoonsa. Tämä prosessi koostuu käytännössä useasta pienestä muunnoksesta, jotka kuitenkin suurpiirteissään voi yleistää seuraaviin pääkohtiin: Alussa yksinkertainen XML-datana oleva sanalista, jossa yhtä sanaa voisi vastata vaikkapa tietue `<st><s>valo</s><t><tn>1</t></st>`. Tämä muunnetaan SFST-leksikkotiedostossa olevaksi sanalistaksi, joka kuvaa identiteettirelaation sanojen perusmuodoista ja taivutusluokituksista, sekä muusta lisädatasta, joka järjestelmälle saattaa olla tarpeellista, kuten epäsäännöllisen vartalovokaalin olemassaolosta. Tämä tekstitiedosto koostuu riveistä, joista kullakin on yksi sana sanakirjamuodossaan, ja muu data kulmasulkein eroteltuna, esimerkiksi sana *valo* tulee muotoon *valo<N1>*

Seuraavaksi sanoista muodostetaan replace-sääntöjen ja konkatenoitavien täydennysten avulla sellaiset taivutusvartalot joita voidaan käyttää taivutuksen lähtökohtana. Nämä taivutusvartalot ovat suunnilleen sellaisia, joihin voi konkatenoida taivutuspäätteitä tekemättä suurempia muutoksia, ja käytännössä lähellä koulukielioppien käyttämiä suomen sanojen taivutusvartaloita (konsonanttivartalo, monikkovartalo jne.) (Setälä, 1930). Esimerkiksi sanalle joka on leksikossa muodossa `käsi<N27>` saadaan taivutusvartalot suunnilleen muotoon `kä<~T>e<N27>`. Lopuksi konkatenoidaan taivutuspäätteet oikeisiin vartaloihin ja toteutetaan äännemuunnokset, jotka kuvaukseen kuuluvat, jolloin esimerkiksi päästään muotoon *kädessä*, joka liittyy toisella tasolla olevaan muotoon `käsi<N27><sg><ine>`. Lopullinen transduktori on kokonaisuudessaan muotoa, jossa on kaikkien sanojen ja niiden taivutusmuotojen vastaavat relaatiot tällaisina yhdistelminä.

Malli, jolla olen tämän prosessin päätynyt tekemään (ks. kuva 2), on lineaarisesti etenevä kuvaus, joka ottaa syötteenään edellisen vaiheen tulosteen, ja muodostaa uuden version äärellistilaisten operaatioiden avulla. Syöte ja tuloste ovat tässä sanojen tai sanamuotojen lista transduktorina, joista periaatteessa joka vaiheessa jokaista sanamuotoa ikään kuin operoidaan. Käytännön tasolla lista tietenkin äärellistilaisessa järjestelmässä on sanamuotolistan disjunktion kanssa ekvivalentti kuvaus, mutta kuvauksen selventämiseksi esitän operaatiot kuin ne operoitaisiin

Kuva 2: Morfologia-järjestelmän kaaviotaulukko

Syöte	**Operaatio**	**Operandi**	**Tuloste**
Nykysuomen XML-sanalista	XSLT	XSLT-kohentimet	Parannettu XML-sanalista
Paranneltu XML-sanalista	XSLT	Leksikkomuunnin	SFST-leksikko
SFST-leksikko	luku	Sanaston luku	Sanalista-transduktori
Sanalista-transduktori	kompositio konkatenaatio	erillissiistintä, typistyssäännöstö, variantit vartalo-osat	 taivutusvartalo-listatransduktori
Taivutusvartalot	konkatenaatio	taivutuspäätteet	taivutusmuoto-listatransduktori
Taivutusmuodot	kompositio	fonologia-toteutussäännöt	valmis sana-listatransduktori

Kuva 3: Äärellistilainen järjestelmän yksinkertaistettu hahmo koodina

```
(($sanalista$ || $taivutusvartaloiksi$)\
  $taivutuspaatteet$)\
  || $fonologia$
```

yksittäisille sanamuodoille kerrallaan. Lisäksi on tietysti niin, että operoinnin sijasta järjestelmä on vain sanalistatransduktorin, sääntötransduktorien ja operaatioiden yhdistelmä, mutta sen hahmottaminen sanamuotojen käsittelynä käsittelynä on mielestäni havainnollista ja perusteltua.

Kuvan 2 taulukkoa luetaan siten, että vasemmassa sarakkeessa on olio, jota operoidaan toisen sarakkeen operaattorilla kolmannen sarakkeen oliota vasten jotta saadaan neljännen sarakkeen olio. Taulukosta on tarkoituksella jätetty pois toteutuksen detaljit, kuten siistinnät, minimoinnit ja filtteröinnit, jotka ovat melko triviaaleja ja muuttuvat järjestelmän eri versioiden välillä paljon; käytännössä turhia merkkejä ja yligeneroivia konkatenaatioita filtteröidään aina kun se on järkevää.

Seikkakohtaisempi kuvaus koko järjestelmän toteutuksesta osa osalta on kappaleessa 3.2.2, ja vielä tarkemman kuvan saa tietenkin tarkastelemalla järjestelmän lähdekoodia, joka on saatavilla osoitteesta `https://gna.org/projects/omorfi`. Järjestelmän toteutus ei nähdäkseni sisällä ei-triviaaleja sääntöjä tai uusia kehityksiä, joten sen selvittäminen lähdekoodista lienee helppoa.

3.2.1 XML-sanalistan muunnos SFST-leksikoksi XSLT-menetelmällä

Lähteenä käytettävä XML-sanalista ei ole sellaista muotoa, joka toimisi SFST-ohjelman syötteenä, joten se on esikäsiteltävä ja muunnettava SFST:n käyttämään leksikkomuotoon, joka on yksinkertainen rivipohjainen tekstiformaatti. XML-lähteen muuntamiseen on tässä päätetty käyttää myöskin XML-pohjaista XSLT-muunnoskieltä, jonka ensisijainen tarkoitus on puurakenteisten dokumenttien kuvaaminen. Tässä tapauksessa kohdemuoto ei ole varsinaisesti niinkään puurakenteinen kuin suora lista, mutta kuvaus toimii silti. Kuvauksen periaatteenahan on ottaa XML-sanalistapuusta jokainen alkio ja lisätä siitä sanakirjamuoto sekä taivutusluokitus sanalistaan. Lisäksi poikkeustapauksista joutuu tekemään pieniä muunnoksia, kuten monta taivutusluokkaa sisältävien sanojen tulostaminen monesti.

XSLT-muunnoskieltä käytetään tutkielman projektissa kahteen tarkoitukseen, sen lisäksi että sillä tehdään muunnos XML-muodosteesta SFST:n ymmärtämäksi lexluetelmaksi, käytettävissä on joukko XSLT-muunnoksia, jotka sisältävät tiettyjä automaattisia korjauksia sanalistaan. Jälkimmäinen kohennusosio ei ole pakollista suorittaa ennen muunnosta lex-listaksi, mutta se parantaa monikkosanojen ja vierasperäisten sanojen käsittelytarkkuutta. Sanalistamuunnoksen koodi on tiedostossa `kotus.xslt`.

Automaattisen kohennuksen suorittama koodi on jaettu tiedostoihin `xmlattribuutit.xslt`, `monikkosanat.xslt`, `monikkotaivutus.xslt` ja `vartalovokaali.xslt`. Kuvauksien ohella on pätkiä koodeista, jota käytetään muunnokseen, mutta täydet koodit kommenteine konteksteissaan löytyvät verkko-osoitteesta `https://gna.org/projects/omorfi`.

XML-attribuutit lisäävä muunnos `xmlattribuutit.xslt` ei tee mitään muuta kuin tarkistaa tiedoston XML-muotovaatimuksia, kuten luonnollisen kielen määrittelevän `@xml:lang`-attribuutin sisältävän suomen kielikoodin, nimiavaruuden osoitteen `@xmlns`-attribuutin, johon olen väliaikaisesti sijoittanut oman osoitteeni `http://www.helsinki.fi/~tapirine/xmlns/experimental/kotus-sanalista`, sekä uniikin tunnisteen `@xml:id`-attribuutin, jolla olen merkinnyt version listasta omakseni ns. id-signature-arvolla `www-helsinki-fi-tapirinen` juurialkiossa. XSLT-koodina tämä muunnos on triviaalisti templaatti, jossa on yllä mainitut attribuutit:

```
...
<xsl:element name="kotus-sanalista"
             namespace="{$kotus-xmlns}">
  <xsl:copy-of select="@*"/>
```

```
  <xsl:attribute name="xml:id"
   namespace="http://www.w3.org/XML/1998/namespace"
   select="$xmlid-signature"/>
  <xsl:attribute name="xml:lang"
   namespace="http://www.w3.org/XML/1998/namespace"
   select="'fi'"/>
...
```

Monikkosanat arvaava käsittelin `monikkosanat.xslt` yrittää arvata sanamuodot, jotka näyttävät monikoilta, eli karkeasti ne, joiden perusmuoto päättyy t-kirjaimeen vaikka sen ei normaalisti kuuluisi, ja merkitsee näihin sanoihin attribuutin `@muoto`, jonka arvona on `NOM PL?`, merkitsemässä arvattua nominatiivin monikkoa. Esimerkiksi sana *aivot* kuten myös sana *aivokuollut* merkittäisiin tässä mahdollisesti monikkosanaksi. Kysymysmerkki merkitseekin sellaista muotoa, joka tulisi tarkistaa ja korjata käsin, ennen kuin sitä käytetään järjestelmässä. Korjaaminen tapahtuu poistamalla kysymysmerkki, jos arvaus on oikein, tai muuttamalla sen muodoksi `NOM SG`, jos kyseessä on yksiköllinen sana. Jos attribuutti poistetaan niin tämä automaattinen tarkistus lisää sen uudelleen seuraavalla tarkistuskerralla, joten poistaminen ei kannata. Lisäksi olen kokeillut attribuuttia `@perusmuoto`, johon arvataan sanan rekonstruoitu perusmuoto, eli sellainen jonka voisi olettaa yksikön nominatiivin olevan, jotta sanan käsittely morfologisessa jäsentimessä sujuisi oikein. Tätä kenttää voisi käyttää vaikkapa silloin, kun morfologisen järjestelmän taivutusosa ei osaa palauttaa monikkomuotoista sanakirjamuotoa oikein käsittelyä varten. Uusista attribuuteista on huomattava, että alustavasti olen sijoittanut ne toiseen nimiavaruuteen kuin muut attribuutit käyttäjille vihjeeksi siitä, että kyseessä on muualta tullutta, epävarmempaa dataa kuin Kotimaisten kielten tutkimuskeskuksen antamaa. Monikkosanojen arvaus käsittellään XSLT 2.0:n `analyze-string`-ominaisuudella jokaista sanan jokaista taivutusmuotoa kohden seuraavasti:

```
<xsl:analyze-string select="s" regex="^(.*)t$">
  <xsl:matching-substring>
    <xsl:call-template name="tap:monikkoattribuutit">
      <xsl:with-param name="yksikkö"
       select="regex-group(1)"/>
    </xsl:call-template>
  </xsl:matching-substring>
</xsl:analyze-string>
```

Monikkotaivutuksen merkitsevä käsittelin `monikkotaivutus.xslt` täydentää monikkosanakäsittelintä sillä, että se merkitsee monikkosanojen taivutuksen koskevan vain sanan monikkomuotoja, eli käyttää `t`-alkion olemassaolevan

`@taivutus`-attribuutin ennalta tunnettua arvoa `monikossa`. Monikkotaivutuksen lisäyksen koodi tarkistaa yksinkertaisesti monikkovaiheessa lisätyn attribuutin:

```
<xsl:if test="../s/@tap:muoto = 'NOM PL'">
  <xsl:attribute name="taivutus" select="'monikossa'"/>
</xsl:if>
```

Vartalovokaalin arvaava käsittelin `vartalovokaali.xslt` on yksinkertainen luokan 22 vierassanojen taivutusta helpottava koodi, joka pyrkii arvaamaan tietyntyyppisten sitaattilainojen äännerakenteen käyttäen pohjaoletuksenaan, että nämä ovat lainasanoja ranskasta tai englannista, kuten alkuperäisessä Nykysuomen sanalistassa kaikille tämän luokan sanoille pätee. Vartalovokaali tarkoittaa tässä sitä sanan lopussa ääntyvää vokaalia, joka ratkaisee vokaalisoinnun ja illatiivin tunnuksen, esim: show : show'ta : show'hun tai bordeaux : bordeaux'ta : bordeaux'hon. Tämä tunnistus on XSLT:ssä yksinkertainen merkkijonontäsmäysfunktio, jota kutsutaan jos sanan taivutusluokka on 22. Tunnistus on lähinnä esimerkinomainen, ja toimii vain sanoille, jotka olivat Nykysuomen sanalistan versiossa 1 mukana, muissa sitaattilainoissa sen arvaukset eivät välttämättä ole oikeita.

```
<xsl:if test="tn = 22">
  <xsl:attribute name="vartalovokaali"
   select="tap:vartalovokaali22(../s)"
   namespace="{$minun-xmlns}"/>
</xsl:if>
...
<xsl:function name="tap:vartalovokaali22" as="xs:string">
  <xsl:param name="sana" as="xs:string"/>
  <xsl:choose>
    <xsl:when test="matches($sana, 'illes$')">i</xsl:when>
    <xsl:when test="matches($sana, 'eaux$')">o</xsl:when>
    <xsl:when test="matches($sana, 'ait$')">e</xsl:when>
    <xsl:when test="matches($sana, 'et$')">e</xsl:when>
    <xsl:when test="matches($sana, 'at$')">a</xsl:when>
    <xsl:when test="matches($sana, 'ut$')">u</xsl:when>
    <xsl:when test="matches($sana, 'w$')">u</xsl:when>
    <xsl:otherwise><xsl:message>
      Tunnistamaton vartalovokaali sanassa
      <xsl:value-of select="$sana"/>
      </xsl:message>?</xsl:otherwise>
  </xsl:choose>
</xsl:function>
```

Itse varsinainen muunnos XML-muodosta SFST-muotoon toimii suurpiirteissään niin, että luetaan `kotus-sanalista`-puusta, eli tämän version juuresta, jokainen lapsi, joka on tyyppiä `st`, eli kaikki sanatietueet nykyisen määritelmän mukaan. Näistä tarkastetaan `t`-alkion eli taivutusdatan olemassaolo, ja viskataan pois ne, joista se puuttuu. Seuraavaksi käydään läpi jokaisen tietueen kaikki `t`-lapset, eli käsitellään saman sanan eri taivutusluokat eri olioina. Näistä `tn`-lasten sisällön perusteella lisätään muunnoksen tulokseen joko nominirivi, sanaluokille 1—49 tai verbirivi sanaluokille 52—78, poimimalla tulosriviin sisällöt käsiteltävän `t`-alkion sisaresta `s` (sanakirjamuotoinen sana) ja lapsesta `tn` (taivutusnumero) sekä `av` (astevaihtelu), jos sellainen on olemassa. Lisäksi jos astevaihtelussa on attribuutti `@astevaihtelu='valinnainen'`, tulostetaan sama sana kahtena rivinä, kerran astevaihtelun kera ja kerran ilman. Jos sanalistaa on kohennettu XSLT-skripteilläni ja monikkosanat merkitty monikkotaivutus-merkillä, lisätään tämä tieto mukaan riville. Kuvassa 4 on otos muunnoksen vaiheista. Koodi, joka toteuttaa muunnoksen, koostuu jokaisen sanatietueen jokaisen taivutustiedon läpikäynnistä ja datojen ylöskirjoituksesta seuraavasti:

```
<xsl:template match="st">
  <xsl:if test="not(t)">
    <xsl:message terminate="no">
      Sanaa <xsl:value-of select="s"/> ei tulostettu
      leksikkoon, koska sen taivutusluokka uupuu.
    </xsl:message>
  </xsl:if>
  <xsl:for-each select="t">
    <xsl:choose>
    <xsl:when test="1 &lt;= tn and tn &lt;= 49">
      <xsl:value-of select="../s"/>
      «N<xsl:value-of select="tn"/>»
      <xsl:if test="av">
        «AV<xsl:value-of select="av"/>»
      </xsl:if>
      <xsl:if test="@taivutus='monikossa'">«PLT»</xsl:if>
    </xsl:when>
    <xsl:when test="51 &lt;= tn and tn &lt;= 78">
      ...
    </xsl:when
    </xsl:choose>
  </xsl:for-each>
</xsl:template>
```

Kuva 4: XSLT-muunnoksen vaiheita

kotus-sanalista_v1.xml:

```
<kotus-sanalista>
<st><s>valo</s><t><tn>1</tn></t></st>
<st><s>valoaallot</s></st>
<st><s>valohämy</s></st>
<st><s>valoisasti</s><t><tn>99</tn></t></st>
<st><s>valottaa</s><t><tn>53</tn><av>C</av></t></st>
<st><s>valvojaiset</s><t><tn>38</tn></t></st>
<st><s>uros</s><t><tn>39</tn></t>
  <t><tn>41</tn></t></st>
```

→kotus-sanalista_v1-r1.xml:

```
<kotus-sanalista
 xmlns="http://www.helsinki.fi...
 xmlns:tap="http://www.helsinki.fi/~tapirine/..."
 xml:id="www-helsinki-fi-tapirine1" xml:lang="fi">
<st><s>valo</s><t><tn>1</tn></t></st>
<st><s tap:muoto="NOM PL?"
 tap:perusmuoto="valoaallo">valoaallot</s></st>
<st><s>valohämy</s></st>
<st><s>valoisasti</s><t><tn>99</tn></t></st>
<st><s>valottaa</s><t><tn>53</tn><av>C</av></t></st>
<st><s tap:muoto="NOM PL?"
 tap:perusmuoto="valvojainen">valvojaiset</s><t
 taivutus="monikossa"><tn>38</tn></t></st>
<st><s>uros</s><t><tn>39</tn></t>
  <t><tn>41</tn></t></st>
```

→kotus-sanalista.lex:

```
valo<N1>
valottaa<N53><AVC>
valvojaiset<N38><PLT>
uros<N39>
uros<N41>
```

3.2.2 Äärellistilaisen järjestelmän konkreettiset SFST-moduulit

Äärellistilainen järjestelmä, joka käsittelee sanalistaa toimii siten, että se lukee sanalistan ulkoisesta datalähteestä, leksikkotiedostosta, joka on kappaleessa 3.2.1 kuvatun XSLT-muunnoksen tulostiedosto. Tästä sanalistasta järjestelmä toteuttaa vartaloissa esiintyvät äännemuutokset, liittää sanoihin taivutuspäätteet ja toteuttaa niissä esiintyvät ja niiden aiheuttamat äännemuutokset. Lopputuloksena saadaan yksi transduktori, joka kuvaa sanojen kaikki taivutusmuodot niiden perusmuotoihin ja taivutusdataan, eli transduktori, joka on ekvivalentti kaikkien sanamuotojen disjunktion parin sanakirjamuotojen ja kuvausten kanssa.

Systemaattinen toteutus, jolla viitetaulukkojeni (1 s. 20 ja 2 s. 24) kautta pääsee perusmuotolistasta tynkä—vartalo—päätemallin kautta lopputulokseen, on sellainen, että aluksi perusmuodosta leikataan vartalon variantti osa taivutusluokkanumeron perusteella. Esimerkiksi nyt sanasta *lapsi* leikataan luokan 29 perusteella pois loppuosa *psi*, jolloin jäljelle jää vaihtelematon osa *la* sekä taivutusluokkanumeron tunnus 29. Tämän jälkeen lisätään täydennykseksi eri vartaloihin täydentävät variantit osat vartalosta, esimerkiksi *la*-tynkään *s* konsonanttivartaloa varten ja *pse* yksikkövartaloa varten. Soveltuviin vartaloihin edelleen lisätään taivutuspäätteet, siis konsonanttivartaloon *las* muun muassa monikon genetiivin tunnus *ten* tai yksikän partitiivin tunnus *ta*. Taivutuspäätteistä osassa käytetään morfofonologisia muotoja, joissa esimerkiksi sointuvokaali *a* tai *ä* on kuvattu yhteisellä merkillä, jonka oikea muoto päätetään seuraavassa vaiheessa. Esimerkiksi muodon *lasta* partitiivin tunnus on muodossa `t<~A>`, ennen kuin vokaalisointu käsitellään. Tarvittavat yleistykset saadaan katsomalla tuleeko joihinkin taivutusluokkanumeroihin samoja sääntöjä tai lisäyksiä. Tässä mallissa vartalot ovat lähellä koulukielioppien kuvauksia sanojen taivutusvartaloista, esimerkiksi nomineille muodostuvat monikkovartalot ja usein konsonanttivartalot ovat täsmälleen odotuksenmukaisia.

Käytännön tasolla äärellistilainen järjestelmä on jaettu aiheittain moduuleihin, ja jokaisessa moduulissa on joukko transduktoreja, jotka toteuttavat yhden konseptuaalisen muunnoksen sanalistalle. Jokainen moduuli on toteutettu siten, että se ottaa syötteekseen sanalistan edellisestä moduulista tulleen version, ja yleensä replace-säännöstön ja kompositio-operaation tai konkatenaatio-operaation ja disjunktioluettelotransduktorin avulla muodostaa uuden sanalistan.

Moduulit on jaettu siten, että jos jonkin osan järjestelmästä voisi toteuttaa toisella lähestymistavalla, olisi se toteutettavissa nimenomaan joitakin moduuleja korvaamalla. Esimerkiksi muunnos sanalistasta taivutusvartaloihin on tässä järjestelmässä kahden moduulin avulla tehty typistys-ja-täydennysoperaatio, mahdollista olisi myös hoitaa sama yhdellä replace-operaatiolla ja korvata nämä kaksi moduulia sillä. Teoriassa jokaisen moduulin transduktori olisi mahdollista toteuttaa

myös sanalistasta riippumattomasti ja kääntää erilliseen tiedostoon, ja kompositoida keskeisesti sanalistan kanssa, mutta tätä koettaessani törmäsin rajoitteisiin SFST-järjestelmässä, joita kuvaan tarkemmin kappaleessa 5. Tämä rajoite tekee järjestelmän muuttelemisen vähemmän triviaaliksi, mutta toteutettevaksi tehtäväksi.

Järjestelmän moduulit ovat suoritusjärjestyksessä `sanat.sfst`, `yksikoksi.sfst`, `paikanna-av.sfst`, `typista.sfst`, `vartaloiksi.sfst`, `taivutus.sfst`, `fonologia.sfst` ja `siisti.sfst`. Lisäksi järjestelmässä on `omorfi.sfst`, joka keräilee lopullisen sanatransduktorin kasaan. Lisäksi järjestelmässä on erillisessä moduulissa `aakkosto.sfst`, joka sisältää apumäärityksiä järjestelmässä käytettyjen merkkien ryhmittelyjen suhteen.

Koodia kuvattaessa olen esittänyt pätkiä koodeista, jotka ovat relevantteja välttäen samalla saman koodin toistoa tai trivialiteettien esittelyä. Täydet ja ajantasaisimmat koodit on helpointa saada luettavakseen osoitteesta `https://gna.org/projects/omorfi`.

Käytännön tasolla moduulit yhdistyvät järjestyksessä toisiinsa kaavalla, joka on yksinkertaistetussa muodossaan kuvassa 5. Kuvan SFST-tyylinen pseudokoodinotaatio kuvaa niitä operaatioita, joita järjestelmän transduktoreihin sovelletaan. Koodin lopussa on havainnollisia esimerkkejä siitä, mitä tyyppiä itse transduktorit aina ovat.

Järjestelmä toimii järjestyksessä siten, että `sanat.sfst` lukee sanalistan identiteettikuvausten disjunktio -transduktoriksi, jossa on sanat muodossa `<wb>sana<tl><av><plt><^x>`, jossa `<wb>` on sanaraja, `sana` on sana sanakirjamuodossaan, `<tl>` taivutusluokan tunnus, `<av>` astevaihtelun tunnus, jos sellainen on, `<plt>` monikkosanatunnus ja `<^x>`[11] poikkeavan vartalovokaalin tai vokaalisoinnun merkki — esimerkiksi *tikkaat*$^{41-A}$ luettaisiin muodossa `<wb>tikkaat<N41><AVA><PLT>`.

`Yksikoksi.sfst` on kompositoitava säännöstö, joka keksii monikkosanojen yksikön nominatiiveja taivutusta varten, esimerkiksi *tikkaat*-sanasta säännöt tekisivät muodon `<wb>tikas<N41><AVA><PLT>`.

`Paikanna-av.sfst` on kompositoitava replace-säännöstö, joka etsii astevaihtelun alaisen äänteen taivutusluokkanumeron ja astevaihtelukirjaimen perusteella ja merkitsee sen astevaihtelevaksi symbolilla `<~k>`, `<~p>` tai `<~t>`, esimerkik-

[11] käytetyissä tunnuksissa ylipäätään on noudatettu sellaista kaavaa, että suuraakkosin kirjoitetaan lähinnä konkatenoinnissa ja filtteröinnissä käytetyt leksikkokoodin kaltaiset merkit, sirkumfleksillä alkavat merkinnät ovat muistiinpanoja, jotka eivät realisoidu vaan esittävät kontekstin ja aaltoviivalla alkavat ovat morfofoneemeja, joilla on useita mahdollisia reaalistumia

Kuva 5: Järjestelmän koodit kaavana

```
 sanat || yksikoksi1 || ··· || yksikoksi24
|| paikanna-av1 || ··· || paikanna-av26
|| typista1 || ··· || typista12

typistetyt vartaloiksi

(vartalot || vartalofiltteri1) |
··· |
(vartalot || vartalofiltteri18)

vartalot taivutukset.

(taivutetut || taivutusfiltteri1) |
··· |
(taivutetut || taivutusfiltteri12) ||
··· ||
taivutusfiltteri18 ||
fonologia1 ||
··· ||
fonologia36.

siisti1 || fonologisoidut || siisti2.
siisti3 || siisti4 || siistityt.
```

, jossa:

`yksikoksi`$_n$ tyyppiä
`(t:<> ^→ (__<NX><PLT>) )`
`paikanna-av`$_n$ tyyppiä
`(k:<k> ^→ (__[^k]$vahva$<AVA>]) )`
`typista`$_n$ tyyppiä
`(i:<> ^→ (__<N5>) )`
`vartaloiksi` on
`<N5>i<VNNOMI> | <N5>e<VNYKSI> | ···`
`vartalofiltteri`$_n$ tyyppiä
`.* <N1> .* <N1> .*`
`taivutukset` on
`<VNYKSI>lle<sg><all> | <VNMONI>ille<pl><all> | ···`
`taivutusfiltteri`$_n$, $n \in [1, 12]$ tyyppiä
`.* <PLT> .* <pl> .* | [^<PLT><pl>]`
`taivutusfiltteri`$_i$, $i \in [13, 18]$ tyyppiä
`.* <VNNOMI> .* <VNNOMI> .*`
`fonologia`$_n$ tyyppiä
`(<n>:l ^→ (l [^#aakkoset#]* __)`
`siisti`$_n$ tyyppiä
`.*`

si *tikkaat*-sanasta säännöt tekisivät <wb>tik<~k>as<N41><AVA><PLT>.

Typista.sfst on kompositoitava replace-säännöstö, joka kuvaa sanaluokittain sanojen vaihtelun alaisen osan vartalosta tyhjäksi ja ottaa muistiin mahdollisesti kadotetun vartalovokaalin, esimerkiksi sanasta *tikkaat* säännöillä tulisi muoto <wb>tik<~k><^a><N41><AVA><PLT>.

Vartaloiksi.sfst on konkatenoitava luetelma niistä vaihteluvarianteista vartalon osista, joista nominatiivia vastaava osa typistettäessä poistettiin, esimerkiksi sanasta *tikkaat* saisi muun muassa monikkovartalon <wb>tik<~k><^a><N41><AVA><~a><VNMONI>. Lisäämällä nämä taivutusluokkakohtaisesti saadaan aikaan kaikki sanan mahdolliset taivutusvartalot, jotka suunnilleen vastaavat Nykysuomen sanalistan mallitaivutustaulukon rivejä pienin muutoksin ja yleistyksin.

Taivutus.sfst on konkatenoitava luetelma kaikista taivutuspäätekombinaatioista, jotka lisätään soveltuviin taivutusvartaloihin, eli sijamuotojen, possessiivien ja kliittien kombinaatioista nomineille ja aika-, tapa ja persoonamuotojen sekä kliittien ja infiniittimuotojen sekä nominaalitaivutusten kombinaatioista verbeille. Sanaan *tikkaat* lisäämällä saisi muun muassa monikon inessiivin <wb>tik<~k><^a><N41><AVA><PLT><~a><pl>i<ine>ss<~a>.

Fonologia.sfst sisältää säännöt niille morfofonologisille yleistyksille, jotka järjestelmässä ovat olleet konkatenaatiokäsittelyn yksinkertaistamiseksi, kuten vokaalisoinnun ratkaiseminen päätteiden tunnuksissa tai astevaihtelun toteutus. Tässä vaiheessa sanan *tikkaat* monikon inessiivi päätyisi muotoon <wb>tikk<^a><N41><AVA><PLT>a<pl>i<ine>ssa.

Siisti.sfst poistaa vielä transduktorista kaikki väliaikaiset symbolit, jota transduktorien rakennuksessa, yhdistelyssä ja suodattamisessa on käytetty apumerkkeinä, kuten leksikkosymbolit ja odotuksenvastaisen vokaalisoinnun merkitsevät symbolit, jolloin sanan *tikkaat* generointitasolle saadaan tikkaissa. Tässä vaiheessa lienee huomion arvoista, että analyysitasoon ei ole koskettukaan paitsi potentiaalisesti konkatenaatio-operaatioissa, joten analyysitasolle jää sanat.sfst:n tikkaat<N41> ja taivutus.sfst:ssä molemmille tasoille konkatenoitu tunnistejono <pl><ine>, eli haluttu analyysi tikkaat<N41><pl><ine>.

Yhteenveto taivutusprosessista on taulukossa 6. Taulukossa on mallisana, joka kuvaa järjestelmässä etenemistä taso tasolta kuten se varsinaisessa järjestelmässä toimii siten, että ylemmällä rivillä on transduktorin analyysitaso ja alemmalla rivillä on transduktorin generaatiotaso. Merkintöjä on yksinkertaistettu siten että vain merkitykselliset merkit näkyvät kullakin tasolla, käytännössä merkit <N41><AVA><PLT> kulkevat mukana aina siistintään asti molemmilla tasoil-

Kuva 6: Äärellistilaisen järjestelmän moduulit ja niiden tuottamat transduktorit järjestyksessä

Moduuli	**Esimerkki**
sanat	`<wb>tikkaat<N41><AVA><PLT>` `<wb>tikkaat<N41><AVA><PLT>`
yksikoksi	`tikkaat<N41><AVA>` `tik as <N41><AVA>`
paikanna-av	`tik k aat<N41><AVA>` `tik <~k>as <N41><AVA>`
typista	`tik k aat <N41>` `tik <~k><^a><N41>`
vartaloiksi	`tikk aat <N41>` `tik<~k><^a><N41><~a><VNMONI> |...`
taivutus	`tikk aat <N41><pl><ine>` `tik<~k><^a><~a>iss<~a><N41><pl><ine> |...`
fonologia	`tikkaat <N41><pl><ine>` `tikkaissa<N41><pl><ine> |...`
siisti	`tikkaat<N41><pl><ine>` `tikkaissa |...`

la, samoin kuin uudet mukaan tulevat merkit, kuten leksikkosymboli `<VNMONI>`, mutta tarkemmin toteutuksesta saa selville itse lähdekoodista.

3.3 Korpusmenetelmät

Testaamiseen käytetty korpusdata on CSC — Tieteellinen laskenta Oy:n ylläpitämässä Suomen tekstipankissa olevaa TEI XML:stä laajennettua XML-merkattua dataa, ja sitä käsitellään testattavaksi samankaltaisilla XSLT-menetelmillä kuin sanalistan koodi. XSLT-käsittelyn tuloksena saatu data käsitellään Python-kielisellä sovelluksella, joka käyttää pysfst[12]-nimistä liitäntää analysoidakseen korpuksen sanamuotoja tekemälläni transduktorilla, sekä verratakseen analyysijärjestelmäni antamia tuloksia tekstipankissa oleviin merkintöihin, jotka eivät ole täysin vastaavia. XSLT-käsittelin on kuvattu lyhyesti kappaleessa 4, ja python-käsittelimen lähdekoodit löytää helpoiten omorfin jakelupaketista tai verkosta `https://gna.org/projects/omorfi`.

[12] `https://gna.org/projects/pysfst`

4 Testiasetelma ja evaluointi

Rakentamastani morfologisesta järjestelmästä testaan kahta asiaa. Ensin selvitän käytetyn muoto-opin kuvauksen toimivuutta ja onnistuneisuutta, sekä sanaslistan kattavuutta testaamalla korpus-aineistoa vasten jäsentimen saantia, tarkkuutta ja kattavuutta. Toiseksi testaan käytetyn SFST-järjestelmän suorituskykyä ajastamalla ja mittaamalla muistinkäyttöä sekä järjestelmän suoritustilanteessa, että järjestelmän rakentaessa transduktoria säännöstöistä ja operaatioista sekä sanalistasta.

Korpustestissä käytin aineistona CSC:n ylläpitämästä kielipankista löytyvää vapaaseen käyttöön tarkoitettua, valmiiksi morfosyntaktisesti jäsennettyä korpusta, jonka morfologisia tulkintoja ei ole käsin tarkistettu. Morfologisen tulkinnan vertausaineistona on siis toisen automaattisen jäsentimen antamia tulkintoja juoksevasta sanomalehtitekstistä.

Suorituskykymittaukset olen tehnyt yksinkertaisesti mittaamalla suorittimen- ja muistinkäyttöä järjestelmästä tavallisilla GNU-työkaluilla. Tällä tavoin olen saanut riittävästi suuntaa antavia tuloksia, joita voi käyttää suorituskyvyn havaitsemisessa ja vertailussa, mutta tarkemmat tutkimukset olen sivuuttanut. Suorituskykymittauksissa vertailukohteena on toisen äärellistilaisen järjestelmän mittaukset olosuhteissa, jotka on pyritty mahdollisimman tarkkaan toisintamaan vastaamaan omaa järjestelmääni SFST:ssä.

Kaikki kappaleessa kuvatut testit on suoritettu CSC — tieteellinen laskenta Oy:n koneessa, joka sijaitsi verkko-osoitteessa `corpus3.csc.fi`. Testauskoneen valinta perustuu siihen, että korpukset on lisensoitu tavoin, joka ei salli käyttää niitä muualla, joten corpus3 on ainoa palvelin jolla testauksen voi suorittaa.

4.1 Korpusmittaukset

Korpusmittauksissa vertaan järjestelmäni antamia analyysejä koneellisesti yhteen kielipankin korpusaineistossa vapaalla B-lisenssillä käytössä olevaan, valmiiksi analysoituun aineistoon, joka on otettu sanomalehti Karjalaisen vuosikerrasta 1992. Se sisältää 3 miljoonaa juoksevaa sanetta, joista 2 716 651 on laskettu mukaan analysoitavaksi. Korpusaineiston esikäsittelyn olen toteuttanut XSLT-ohjelmointikielellä, sillä esivalmisteltu korpusaineisto on laajennetussa TEI.2 XML -muodossa. Esikäsittelyni koostuu ainoastaan sanojen poiminnasta analyysiä varten siten, että jos sanan luokaksi on merkitty substantiivi, adjektiivi tai verbi, se lisätään analysoitavaksi. Koodi joka tuottaa analyysirivit XML-lähteestä on seuraava:

```
<xsl:template match="s">
```

```
  <xsl:for-each select="w">
    <xsl:if test="@type='Noun' or
                  @type='Verb' or
                  @type='Adjective'">
      <xsl:value-of select="."/>,
      <xsl:value-of select="@lemma"/>,
      <xsl:value-of select="@msd"/>
      <xsl:text>&#xA;</xsl:text>
    </xsl:if>
  </xsl:for-each>
</xsl:template>
```

Näin suoritetulla koodilla saadaan aineistosta tulokseksi yksinkertainen csv-tyyppinen tiedosto, jossa jokaisella rivillä on kolme pilkuin erotettua kenttää, jotka sisältävät järjestyksessä saneen, sanan sanakirjamuodon, ja morfologiset merkinnät. Tässä tiedostossa jokainen yksittäinen sanamuoto saattaa esiintyä mielivaltaisen monta kertaa, ja yleiset sanat ja sanamuodot vievätkin suuren määrän esiintymistä. Tästä aiheutuvan virhetulkinnan kiertämiseksi suoritin myös toisen testauksen, johon jokainen sanamuoto otetaan mukaan vain kerran. Tästä tiedostomuodosta tällainen aineisto, jota nimitän tyypittäiseksi sanalistaksi, tuotetaan komennolla `sort | uniq`; tässä sanalistassa jokaisen sanan esiintymä on ainoalaatuinen, eli kahta saman sanan samamuotoista analyysiä ei esiinny kahdesti.

Automaattinen analyysi on suoritettu python-kielisellä sovelluksella pysfst-moduulia käyttäen. Analyysi toimii siten, että jokaisesta sanamuodosta tuotetaan järjestelmällä kaikki tulkinnat, ja niistä tulkinnoista, joiden perusmuodot ovat samat, yritetään yhdistää kaikki kielipankissa olevat morfologiset merkinnät toteuttamani järjestelmän vastaaviin. Koska kielipankin aineiston tulkinnat eivät aivan täsmää järjestelmässä käytettyihin, olen pyrkinyt olettamaan niitä samaan suuntaan, siten, että esim. genetiivin tai nominatiivin possessiivin vaihtelut hyväksytään aina yhtenäisinä — itse merkitsen ne kahtena eri analyysinä ja aineisto käyttää erillistä genom-symbolia. Tarkka suoritettu testauskoodi löytyy osoitteesta `https://gna.org/projects/omorfi`.

Korpusmittaukset on suoritettu järjestelmän sellaisella versiolla, johon on otettu mukaan vapaasti kliittejä sanojen loppuun mielivaltainen määrä, sekä partisiippimuotojen ja tekemisjohdoksen eli vanhan IV infinitiivin vapaa produktio ja taivutus. Yhdyssanajohdosta mukana on hyvin karkea nominipareja yhdistelevä versio. Se miten tämä vaikuttaa mittaustuloksiin on kuvattu seikkaperäisemmin kappaleessa 5.4.

Seuraavissa kaavoissa merkitsen sanoiksi jokaista yksittäistä sanamuotoa ja analyyseiksi jokaista erillistä analyysia, joita minun järjestelmässäni tulee 1—n sanaa

kohti ja korpuksessa on 1 per sana. Kielipankin tuntemiksi analyyseiksi niitä, jotka löytyvät korpuksesta ja joille on annettu analyysi, jonka olen ottanut mittaukseen. Oikea analyysi on sellainen, jossa minun järjestelmäni on tuottanut oikean perusmuodon, ja morfologiset analyysit ovat samat tai vastaavat kuin korpuksessa, kaikki muut analyysit ovat vääriä. Nykysuomen sanalistan sanoiksi lasketaan sellaiset, joista kielipankin korpuksen analyysin mukainen perusmuoto löytyy Nykysuomen sanalistasta.

Tulosten kattavuus aineiston yli on laskettu yksinkertaisesti kaavalla

$$\text{Kattavuus} = \frac{\text{Järjestelmäni oikeat analyysit}}{\text{Kielipankin analysoidut sanat}}, \tag{1}$$

joka kertoo miten suuren osan korpuksen kaikista analysoiduista, mukaan lasketuista sanoista järjestelmästä löytyy oikeina tulkintoina.

Toiset korpusanalyysitestit on suoritettu poimimalla korpuksesta niiden sanojen muodot, jotka löytyvät myös Nykysuomen sanalistasta, ja testattu niiden analyysien kannalta kuinka tarkka jäsennin on ja miten paljon siinä on vielä puutteita tai virheitä. Tavoitearvona oli tietysti, että järjestelmä tunnistaa siinä olevien sanojen sanamuodoista kaikki 100 %, mutta systemaattisten puutteiden lisäksi molempien järjestelmien virheet ja eri analyysit tietysti laskevat tätä arvoa.

$$\text{Saanti} = \frac{\text{Järjestelmäni oikeat analyysit}}{\text{Nykysuomen sanalistan sanat} \cap \text{Kielipankin analysoidut sanat}}, \tag{2}$$

joka kertoo miten suuri osa niistä analyyseistä, joiden pitäisi onnistua, menee oikein.

Järjestelmän tarkkuus taas on arvo, joka kertoo, että miten suuri osa annetuista analyyseistä menee oikein niiden sanojen kohdalla, joista tulee useita analyysejä. Koska vertailukohtana tässä on disambiguoiva järjestelmä ja oma järjestelmäni antaa kaikki mahdolliset tulkinnat, kuvaa tarkkuus-tulos lähinnä suomen kielen taivutusopissa esiintyvää ambiguiteettipotentiaalia, ja saatu tulos on siis odotuksenmukainen.

$$\text{Tarkkuus} = \frac{\text{Järjestelmäni oikeat analyysit}}{\text{Järjestelmäni kaikki analyysit}} \tag{3}$$

Kuva 7: Järjestelmän analyysien kattavuus

Testi→ ↓ **Korpus**	**Kattavuus**	**Saanti**	**Tarkkuus**
Karjalainen 1992 2 716 651 sanamuotoa	76,8 %	97,0 %	54,0 %
367 242 tyyppiä	50,1 %	96,2 %	47,0 %

4.2 Suorituskykymittaukset

Suorituskykymittaukseni perustuvat yksinkertaisiin muistin ja prosessoriajan käytön mittauksiin järjestelmää rakennettaessa ja käytettäessä. Mittaustyökaluina toimivat GNU-järjestelmistä löytyvät muistin- ja ajankäytön mittaustyökalukomennot `time` ja `top`. Mitatut ajat ovat timen mukaisia käyttäjän kuluttamia prosessoriaikoja, eivätkä varsinaisia suoritusaikoja. Todellisuudessa esim. FSM-kääntö on huomattavasti hitaampi kuin SFST-kääntö ja SFST:n analyysi optimoimattomalla transduktorilla kestää kertaluokittain enemmän kuin optimoidulla. Laskettaessa suorituskykyä järjestelmästä käytettiin sellaista versiota, joista otettiin mukaan vain perusmuotojen taivutukset, eli substantiiveista sijamuodot ja possessiivisuffiksit sekä verbeistä persoona- ja tapamudot sekä infiniittimuodoista perusmuodot; kaikenkaikkiaan 7 541 661 sanamuotoa. Mukana ei ollut yhdyssanamekanismia, verbin nominaalimuotojen edelleentaivutusta tai arvainta. Versiossa, joka muodostaa tämän on SFST:n tulostiedostossa 25 371 tilaa ja 60 389 kaarta. Järjestelmän aakkoston koko on 456 merkkiä, johon sisältyvät suomen kielessä käytettävät aakkosmerkit, joitakin lainasanoissa käytettyjä latinalaisten aakkosten merkkejä, taivutusluokkien ja muotojen analyysimerkit sekä järjestelmän sisäisiä väliaikaisia analyysimerkkejä, kuten astevaihtelun merkinnät. SFST:stä on käytetty mittaustesteissä versiota 1.1 ja AT&T:n FSMlibistä versiota 4.0. Mittaustuloksissa on vähintään kolmen eri mittauksen keskiarvo; tuloksia mitatessa ei kuitenkaan eri mittauskertojen välillä ollut merkittävää hajontaa.

AT&T:n FSMlibin muodostama transduktori on kooltaan 29 250 tilaa ja 463 121 kaarta ja käytetty aakkosto sama kuin mikä SFST:ssä. Aakkoston osalta huomattavaa kuitenkin on, että FSM määrittelee aakkoston erikseen generointi- ja analyysitasolle, kun taas SFST määrittelee aakkoston symboliparien aakkostona, mikä selittänee ainakin osan siitä miksi samalla menetelmällä tuotetut lopulliset transduktorit järjestelmien välillä ovat merkittävästi eri kokoiset.

AT&T:n FSMlib-järjestelmän testaaminen on suoritettu ottamalla SFST-järjestelmästä erilliset transduktorit irti »-operaattorilla, joka kirjoittaa käsiteltävän transduktorin binääritiedostoksi, ja sen jälkeen muuntamalla SFST:n automaatti `fst-print`-komennolla erääseen tekstimuotoon ja muuntamalla

siitä skriptein FSM:n tekstimuotoon. Tästä lukemalla `fsmcompile`lla saadaan FSM:n binäärimuoto. Mitatussa ajassa on mukana se `fsmcompose`-, `fsmconcat`- ja `fsmunion`-operaatioiden summa, jotka korvaavat SFST-järjestelmässä olleita `||`-, konkatenaatio- ja `&`-operaatiota, vastaavasti. Näiden operaatioiden lisäksi käytin `fsmrmepsilon`- ja `fsmdeterminize`-käskyjä vastaavasti kuin SFST-järjestelmässä, eli joka askeleen jälkeen, ja lisäksi transduktorien muotoina oli `input_indexed` tai `output_indexed`. Lopullisen transduktorin pyrin lisäksi optimoimaan `fsmencode`- ja `fsmminimize`-komennoilla.

Vastaavasti SFST:n käännön nopeustestaaminen on suoritettu samoilla, valmiiksi determinisoiduilla ja minimoiduilla transduktoreilla. Jäljelle jääneiden erojen merkitystä ja syitä pohdin tarkemmin kappaleessa 5.

Sekä AT&T:n FSM että SFST tukevat erilaisia transduktoriformaatteja, eli käytännössä eri tavalla optimoituun muotoon kirjoitettuja transduktoribinäärejä. Testeissä SFST:ltä on kokeiltu kaikkia kolmeja formaatteja siten, että lopullinen transduktori on tuotettu `fst-compiler-utf8 -c`- tai `fst-compiler-utf8 -l`-komenolla sarakkeen mukaan. Näistä -c tuottaa kompaktoidun transduktorin, mutta lisäksi muuttaa läpikäyntialgoritmia tavallisesta backtrack-algoritimiksi (Schmid, 2007a). Komennolla -l tuloksena on lowmem-transduktori, joka on myös kooltaan pieni, ja lisäksi sitä ei lueta käsiteltäessä muistiin vaan käytetään mahdollisimman paljon dataa suoraan kovalevyltä.

Vastaavasti FSM-testi on tuotettu valikoimalla transduktorien muodoksi InputIndexed tai OutputIndexed käsin parhaan lopputuloksen takaamiseksi. SFST:n transduktoreista optimoituja versioita ei voi käyttää lainkaan muuhun kuin analyysiin.

Generointitestin komento SFST:llä on `fst-generate kotus.sfsta > /dev/null`, ja korpusanalyysitestille `fst-infl kotus.sfsta karjalainen1992.wordlist > /dev/null`, missä `karjalainen1992.wordlist` on karjalaisesta kappaleessa 4.1 kuvatulla tavalla irrotetusta sanalistasta 10000 ensimmäistä tietuetta[13]. FSM:llä analyysitesti on tehty `farcompilestring`-komennolla ja indexed-tyypillä, sekä kompositoimalla tulos farfilterillä, ja generointi vastaavasti.

[13]Koko korpuksen analysointi transduktorilla, jonka käännössä ei ole käytetty asetusta -c tai -l kestää n. 24 tuntia

Kuva 8: Järjestelmien suoritinajan ja muistin käytön vertailutaulukko

Suoritinaika				
Järjestelmä→ **↓Toiminto**	SFST	SFST -c	SFST -l	AT&T FSMlib
Kääntäminen	309 s	309 s	309 s	125 s
Generointi (kaikki)	17 s	—	—	9 s
Analyysi (korpus)	251 s	0,3 s	10 s	1,3 s
Muistinkäyttö				
Kääntäminen	55 MiB	55 MiB	55 MiB	83 MiB
Generointi (kaikki)	2,8 MiB	—	—	1,4 MiB
Analyysi (korpus)	2,8 MiB	1,8 MiB	1,1 MiB	1,4 MiB

5 Keskustelu

Vapaan ja avoimen morfologisen järjestelmän kehittäminen vapaista komponenteista lienee ensimmäisiä laajamittaisia suomen kielen kohdalla, vaikka aiempiakin kuvauksia muoto-opista on jo varhainkin (Koskenniemi, 1983). Muissa kielissä tällaisia järjestelmiä on toteutettu jonkin verran, esimerkiksi unkarin kohdalla Trón et al. (2005). Tekemäni järjestelmä on uusi ja varhaisessa vaiheessa, joten seuraavassa kappaleessa yritän kuvata kaikkea mahdollista jatkotutkimusta.

5.1 Jatkotutkimuksesta

Nykyinen järjestelmä generoi kieliopin ja sanalistan kuvauksen mukaisesti kaikista sanoista kaikki muodot allomorfeineen riippumatta siitä ovatko jotkin muodoista ja allomorfeista harvakäyttöisiä tai jopa kuolleita. Tällaisessa testaussovelluksessa se ei tietenkään ole ongelma, mutta jos morfologiaa käyttää käytännön sovelluksissa asia lienee aihetta huomioida. Sama pätee myös järjestelmän vapaaseen yhdyssanamuodostukseen, joka tuottaa huomattavan määrän käyttämättömiä muodosteita, jotka tapauksittain häirinnevät varsinaista sovellusta. Esimerkiksi nominit *utu* ja *ilta* yhdyssanan jälkiosina tuottavat sellaisia sanoja, jotka peittävät olemassaolevia muodosteita verbijohtimelta UtU ja taivutusmuodolta monikon ablatiivi.

Vastaavasti nyt morfologisena analysaattorina järjestelmä pyrkii tunnistamaan ja muodostamaan kaikista sanoista kaikki sanamuodot, joka on käytännöllistä silloin, kun pyritään mahdollisimman kattavasti saamaan tietoa analysoitavan datan morfologisesta luonteesta, mutta muilla sovelluksilla voi olla muunlaisia tavoitteita. Esimerkiksi tekstin oikaisuluvussa voi olla hyödyllistä pitää tiettyjä sanamuo-

toja tai tunnusten allomorfeja harvinaisina tai epäsuotavina.

Ratkaisuksi tässä on mahdollista käyttää sana- ja muotokohtaisia rajoitteita, mutta nykyisillä järjestelmillä on mahdollista kehittää myös sellaisia äärellistilaisia transduktoreja, joihin on liitetty painoja tai todennäköisyyksiä. Eräs tärkeä jatkotutkimuksen aihe lienee, jos tällaista järjestelmää rakentaa, todennäköisyyksien ja painojen perusteiden laskeminen.

Toteutettu järjestelmä generoi sanamuodot käyttäen vahvasti hyväkseen Nykysuomen sanalistan taivutusluokitukseen sisältyvää tietoa, esimerkiksi astevaihtelu, allomorfivalinta ja vartalovaihtelu ovat kaikki toteutettu hyödyntämällä taivutusluokkanumeroita ja astevaihtelukirjaimia eksplisiittisenä datana, vaikka suomen kielen muoto-opin taivutuksesta valtaosa on fonologisesti motivoitunutta, ja siten toteutettavissa myös vaikkapa käyttämällä sanan rakennetta näiden taivutusluokkien asemesta, joka saattaa olla sovelluksittain sopivampi ratkaisu. Toteutuksesta ks. esim. (Koskenniemi, 1983).

Nykysuomen sanalistan sanoista käytin järjestelmässä vain reilusti alle puolia, käyttämättä jäivät luokittelemattomat yhdyssanat, luokitellut yhdyssanat, adverbit ja pronominit. Näistä pronominien käsittely ja osiltaan adverbien vaatisi lähinnä pieniä lisäyksiä järjestelmään. Yhdyssanojen käsittely on kuitenkin vielä ratkaistava asia, samoin kuin adjektiivisuuden tunnistaminen ja adjektiivijohdosten tekeminen. Nimitän tässä adjektiiveja johdoksiksi, sillä morfologisen järjestelmäni kannalta on niin, että komparatiivin ja superlatiivin — sekä tavallansa positiivin — tunnukset ovat sellaisia, joiden tuotokset on lisättävä takaisin säännöstökäsittelyn alkukohtaan, jotta niistä saadaan kaikki tarvittavat taivutusmuodot ulos.

Vastaavasti on auki sanojen epäproduktiivisemman morfologian soveltaminen, eli sellaisten johdostapojen päätteleminen, jotka eivät systemaattisesti käy edes teoriassa sataan prosenttiin morfofonologisista kannoista, vaan rajoittuvat esim. semanttisleksikaalisin perustein. Myös erisnimet vaativat omanlaisensa käsittelytavan, johon järjestelmässä ei ole otettu kantaa.

Nykyinen järjestelmä myös tuottaa jokaiselle sanalle kaikki teoreettisesti mahdolliset tulkinnat, joka on käytännöllistä joihinkin tarkoituksiin, mutta esimerkiksi jos järjestelmää käyttäisi kokonaisten tekstikorpusten morfologiseen merkkaamiseen tarvitsisi kehittää disambiguaatiomenetelmiä, joilla saisi vain lausekontekstissaan oikean tulkinnan aikaiseksi. Disambiguaatiokeinot ovat melko laajasti tutkittu alue, esimerkiksi morfosyntaktisen disambiguaation osalta (Voutilainen, 1997; Karlsson et al., 1995) ja sitten semanttisen (Lindén, 2005).

Järjestelmän laadun ja suorituskyvyn evaluoinnin osalta tarkemmat ja kattavammat kuvaukset olisivat tarpeen (esim. Kanthak ja Ney (2004)). Tutkielmassa esitetyt mittaustulokset ovat kuitenkin parhaimmillaankin yksittäisiä ja suuntaa-

antavia, suorituskykymittaukset on esimerkiksi suoritettu käyttöympäristössä jonka tasalaatuisuutta ei ole voinut taata. Korpustestit on vastaavasti suoritettu vain yhdellä satunnaisesti valitulla korpuksella.

Sanalistan ylläpito ja laajentaminen ovat myös tarpeen. Kuten testaustuloksista selvisi ei sanalistan kattavuus juoksevan tekstin yli kuitenkaan ole vielä kovin suuri. Sanalistan ylläpitoa on tutkittu esimerkiksi pro gradu -tutkielmassa Seppälä (2006).

Eräs puute mikä SFST:tä käyttäessä on huomattavissa on varsinaisen leksikkokäsittelimen puute. Siinä missä esim. tutkielmassa Koskenniemi (1983) ja kirjassa Beesley ja Karttunen (2004) kuvatuissa järjestelyissä sanalistojen ja taivutusmuotojen lisäys hoituu erillisellä leksikkokäsittelyllä, toteuttamassani järjestelmässä tämä vaihe on hoidettu äärellistilaisilla operaatioilla yhdistämällä yligeneroiva konkatenaatio kompositoitavaan filtteriin.

5.2 Nykysuomen sanalistan käytännöllisyydestä

Sanalistassa käytetty XML-formaatti on tällä hetkellä yksinkertainen, mahdollisesti kertakäyttöinen datarakenne, joka ajaa hyvin asiansa. Jos kuitenkin formaattia ajattelisi pitkäaikaisena säilytysformaattina samankaltaisille sovelluksille, se voisi hyötyä muutamista lisäyksistä, jotta se olisi itsenäisenä riittävä. Ensinnäkin käsiteltävien yhdyssanojen kannalta yhdyssanarajan merkintä, eritoten kaikista sanarajoista taipuvien yhdyssanojen osalta olisi helpottava lisä. Astevaihtelun luokituksessa jokainen luokka kuvaa tällä hetkellä vain tyypin. Kuvaamalla myös suunnan olisi sanasta valmiiksi käsiteltävissä jo tarvittava tieto astevaihtelusta, vaikka toki jos suomen kielen morfofonologia oletetaan tunnetuksi jo pelkkä tieto astevaihtelullisuudesta tai -vaihteluttomuudesta riittää oikean luokan ja suunnan arvaamiseksi. Sitaattilainojen taivutusluokan osalta oikea taivutus vaatisi tiedon foneettisesta asusta, joka puuttuu kokonaan, eikä ole juurikaan arvattavissa. Monikkosanoista olisi hyvä tietää että sanakirjamuotona on käytetty listassa monikkoa, jotta sitä ei tarvitsisi arvata. Myös sanan osat olisi hyödyllistä merkata erikseen, vaikka ne sananalkuisista yhdysmerkeistä tunnistaakin. Sanalistassa ei myöskään ole varsinaisesti merkitty adjektiiveja erikseen, vaan vain osa kuuluu tiettyihin taivutusluokkiin yksinomaisesti, mutta osa jakaa taivutusluokkansa substantiivien kanssa. Tämän perusteella komparaatiotaivutus morfologiatoteutuksesta puuttuu, vaikka sen tekeminen ei sinänsä monimutkainen ole.

Lisäksi osa sanoista on joko väärässä taivutusluokassa, tai taipuu muutoin odotuksenvastaisesti, joten pitää olla keino sekä sivuuttaa ne, että antaa korvaava vastine. Näitä sanoja ovat nähdäkseni ainakin numeraalit ja jotkin pronominit, nominit *veri*, *meri*, *aika* ja *poika* yhdyssanamuotoineen sekä *olla*-verbi. Yleistäen

voisi sanoa sanakirjan osalta, että kaikki sanat, joissa kielitoimiston sanakirjassa on sana-artikkelin proosassa selostettu taivutuksen poikkeuksista, tulisi luokitella omiin luokkiinsa sen sijaan että niitä kohdeltaisiin erityisinä.

Toinen morfologisen jäsentimen kannalta mahdollisesti kiinnostava kehitys taivutusluokkanumeroihin on niiden yhdisteleminen yleistysten pohjalta. Triviaalisti on yhdistettävissä ainakin sellaiset kuvaukset, kuten taivutusluokat 24 ja 26, jotka kuvaavat täsmälleen saman taivutuksen, jolla saattaa olla yhdessä allomorfiparissa eri keskinäinen yleisyysjärjestys. Laajemmin, kuten kappaleessa 2 selvitettiin, oikeasti luokittelu kuvaa morfofonologisten piirteiden kombinaatioita, ja usein sellaisia, jotka selviävät sanamuotoja tarkastelemalla, joten siltä pohjalta yleistysten tekeminen voisi johtaa tehokkaampaan järjestelmään.

Myöskin vaikka sanalista onkin laaja, ei se tietenkään käytössä olevan kielen tapauksessa voi olla kattava. Yksi kiinnostava ongelma lieneekin tämän järjestelmän laajentaminen sellaisiin sanoihin joita listassa ei ole. Tekemästäni sanalistaluokitteluselvityksestä saattaa olla hyötyä uusien sanojen luokittelussa ja tuonnissa.

Lisäksi voisi laajentaa järjestelmää säännöllisestä taivutuksesta myös johto-opin morfologian puolelle, millä saisi kiinni uusia ja tulevia ja muita väliaikaisia sanoja, jotka sanalistasta uupuvat.

Näitä XML-muodossa arvioimiani puutteita olen testinomaisesti koettanut järjestelmässäni korjatakin. Monikkosanoja varten lisäsin XML-rakenteeseen attribuutin sanakirjassa olevan sanan muodolle (`s@muoto`), joka monikkosanoilla on monikon nominatiivi (`PL NOM`), ja attribuutin rekonstruoidulle perusmuodolle (`s@perusmuoto`), joka kuvaa sanasta sellaisen muodon, mikä olisi sanakirjan tyypillisen perusmuodon mukainen muoto eli tässä yksikön nominatiivi. Monissa tapauksissahan monikkosanan yksikkömuodot esiintyvät vain yhdyssanan osina tai eivät esiinny kielessä laisinkaan, mutta tässä tapauksessa järjestelmä voi silti käyttää rekonstruoitua yksikköä taivutuksen lähtökohtana, ja poistaa yksikkömuodot lopuksi.

Poikkeuksellisesti taipuvien sitaattilainojen taivutusta varten kokeilin attribuuttia `t@vartalovokaali`, johon tulee sisällöksi sanan ääntöasun mukainen vartalovokaali silloin kun se poikkeaa kirjoitusasun mukaisesta, eli luokan 22 sanoilla aina (esimerkiksi *show'lla u* ja *parfait'lla e*).

5.3 SFST:llä toteutetun suomen kielen morfologian suorituskyky

SFST:n suorituskyvyn tarkastelemiseksi suoritin joitain alkeellisia mittauksia, joiden tulokset on kuvattu kappaleessa 4.2. Lopulta SFST-järjestelmä kääntyy

hitaammin kuin FSM-järjestelmä, mutta suorituskyky valmiilla transduktorilla on hyvä jos käyttää oikeita optimointiasetuksia. Erityisesti on huomattava, että compact-tyypin transduktori analysoi dataa eri algoritmilla kuin muut, ja on useita kertaluokkia nopeampi.

Suorituskyvyn tarkempi evaluointi ja parantaminen vaatii vielä systemaattisempaa ja järjestellympää tarkastelua ja tutkimusta.

5.4 SFST:lla toteutetun suomen kielen morfologian laadusta

SFST-järjestelmän laadullista tasoa mittasin kappaleessa 4.1. Järjestelmä toimi n. 97 % saannilla — verrattuna aiempaan morfosyntaktisesti jäsentävään ja disambiguoivaan järjestelmään — niillä sanoilla, jotka sen oli tarkoitus tuntea. Jäljelle jäävän 3 %:n jakauma on kuvattu taulukossa 9. Taulukossa on järjestetty eroavat analyysit yleisyyden mukaiseen järjestykseen. Merkittävä osa, liki kolmannes on sellaisia on-verbin analyyseja, joita järjestelmäni ei tee laisinkaan, eli passiiviksi merkittyjä muotoja, jotka näyttävät yksikön kolmannen muodoilta. Tämän kuvittelen johtuvan siitä, että vertailujärjestelmä käyttää syntaktista jäsennintä, joka poimii passiivisen tulkinnan muualta. Toiseksi eniten on superlatiiveja, jotka järjestelmästäni oli jätetty tarkoituksellisesti pois, koska kuten aiemmin kuvataan, sanalista ei erottele adjektiiveja substantiiveista. Komparatiivien pienempi lukumäärä johtuu siitä, että mitatussa järjestelmässä oli alkeellinen komparatiivin muodostus lisäämällä mpi yksikkövartaloon. 17 % oli MA-infinitiiviksi merkittyjä agenttipartisiippeja, eli sellaisiksi merkittyjä MA-infinitiivin muotoja, joita MA-infinitiivi ei saa, kuten partitiivia tai genetiiviakkusatiivia (Hakulinen et al., 2004). Kieltoverbiä *ei* ei ole sanalistassa, joten sen muodostus on järjestelmässä osin vajaata, osin sen virhetulkinnat ovat samanlaisia passiiveiksi merkittyjä persoonataivutettuja muotoja kuin olla-sanalla. Sanat *aika* ja *poika* on Nykysuomen sanalistassa merkitty säännölliseen taivutukseen, vaikka niillä on säännöistä poikkeava *i : j* -vaihtelu, jota en ole korjannut. Korpuksessa on myös merkittävä osa johdoksia, joiden kantasana on arvattu aika kaukaa, mitä järjestelmäni ei tee. Näistä merkittävät 3 % on kieltopartisiipiksi nykyään nimettyyn *mAtOn*-johdostyyppiin, eli ma-infinitiivin karitiiviin, kuuluvia. Tarkistamattomia monikkosanoja, jotka sanalistasta löytyvät, mutta joiden monikkous on käsin tarkastamatta, on jopa vajaa prosentti virheistä. Minen-johdoksen tulkintavirheistä kaikki kuuluvat mis-loppuisiin yhdyssanamuotoihin, joita jostain syystä on korpuksesta löytynyt itsenäisinä. Lopuissa noin kolmessa prosentissa virhetulkintoja lähinnä yksittäisiä virhetyyppejä, esimerkiksi kirjoitus- tai ajatusvirheellisiä muotoja, tyyppiä *oikeaseen* illatiivina, tai väärää vokaalisointua, tyyppiä *analyyseissa*. Paljon on myös monikoita, jotka on merkitty yksiköiksi, kuten *alueiden*.

Kuva 9: Järjestelmäni ja korpusdatan analyysien erot Karjalainen 1992 -korpuksessa

Eron tyyppi	**N**	**%**
Olla-passiivit	19056	29,3 %
Superlatiivit	14103	21,7 %
IIIinfiksi merkitty partisiippi	10945	16,9 %
Ei-muodot	6728	10,4 %
Komparatiivit	3024	4,7 %
Muut olla-muodot	2770	4,3 %
Aika	2629	4,1 %
IIIinfiksi merkitty karitiivi	1794	2,8 %
Monikkosanat	600	0,9 %
Poika	501	0,8 %
Pass 3P	185	0,3 %
Minen-johdokset	169	0,3 %
Mon.gen -in	159	0,2 %
Kons.vartalot	99	0,1 %
Iparticiksi merkitty johdos	86	0,1 %
Muita	2076	3,2 %
Yhteensä	64922	100 %

Järjestelmäni kattavuudeksi sain 75 %, tämä tarkoittaa, että korpuksen kaikista saneista, jotka korpusta tehdessä käytetty järjestelmä tunnisti, minun järjestelmäni tunnisti oikein kolme neljännestä. Jäljelle jäävään neljännekseen kuuluu edellä mainitun 3 %:n lisäksi yhdyssanoja, nimiä ja muita vierassanoja. Näistä olen poiminut 1000 sanan otoksesta esimerkiksi taulukon 10. Taulukossa yhdyssanoiksi on merkitty kaikki sellaiset sanojen sanalistassa olevien perusmuotojen yhdistelmät, jotka ovat jääneet tunnistumatta, joissa ei ole sellaista johdosainesta, jota järjestelmäni ei käsitellyt. Vastaavasti johdoksiksi on merkitty paitsi yksisanaiset johdostyypit, joita ei ollut sanalistassa, kuten *UUs*-loppuiset ominaisuuden johdokset, myös yhdyssanat, joiden osassa on esimerkiksi *jA*-johtiminen tekijänjohdos. Tästä esimerkiksi *palkinnonsaaja* on järjestelmälle vieras johdos, koska *saaja* ei ole sanakirjasana vaan johdos mutta *taantuma-aika* on yhdyssana, koska *taantua* on sanakirjassa ja *taantuma* on järjestelmän käsittämä säännöllinen taivutusmuoto (so. agenttipartisiippi ja ma-infinitiivi). Kirjoitusvirheet sisältävät sellaisia tunnistumattomia sanamuotoja, jotka voisi lauseyhteyden perusteella päätellä selväksi kirjoitusvirheeksi ja joiden osoittamaa sanaa ei löydy sanakirjasta. Etuliitteisiin kuuluvat sanat, joissa alkuosa on tyyppiä *ylä-*, *ala-*, *etu-*, *läpi-* jne.

Kuva 10: Järjestelmäni Karjalainen 1992 -korpuksen puuttuvat sanatyypit (1000:n otoksesta)

Puutteen tyyppi	**N**	**%**
Yhdyssanat	858	85,8 %
Johdokset	70	7,0 %
Puuttuvat	31	3,1 %
Etuliitteelliset	27	2,7 %
Tulkintavirheelliset	5	0,5 %
Komparatiivit	3	0,3 %
Kirjoitusvirhe	3	0,3 %
Superlatiivit	2	0,2 %
Monisanaiset	1	0,1 %
Yhteensä	1000	100 %

Järjestelmän tarkkuus oli 50 %, mikä tarkoittaa käytännössä sitä, että keskimäärin yhtä sanetta vastaisi yksi oikea ja yksi väärä analyysi. Suurelta osin tulos selittyy jo sillä, että suomen kielessä esiintyy monimerkityksisyyttä morfofonologisten analyysien suhteen siinä, että yhtä sanetta voi vastata usean sanan usea taivutusmuoto. Lisäksi on joitain järjestelmässäni olevia systemaattisia monitulkintaisuuksia, jotka pienentävät tarkkuutta, kun verrataan disambiguoituihin tuloksiin. Esimerkiksi nominin n-sijasta luodaan lähes poikkeuksetta genetiivin ja akkusatiivin tulkinta, samoin verbien partisiippimuodoista tulee aina sekä verbiä että partisiippia vastaavan sanan sanamuodon tulkinnat.

6 Yhteenveto

Tutkielmassani olen esitellyt suomen kielen äärellistilaisen automaattisen jäsentimen kehitystä ja testausta. Tutkielman lähtäkohtana oli rakentaa avointa ja vapaata lähdekoodia oleva kokonaisjärjestelmä. Tähän tarkoitukseen sopiviksi resursseiksi valittiin SFST-niminen äärellistilaisten transduktorien ohjelmointijärjestelmä, ja sanalähteeksi Kotimaisten kielten tutkimuskeskuksen julkaisema Nykysuomen sanalista.

Nykysuomen sanalista on taivutustyypeittäin luokiteltu 78 eri luokkaan, joita jäsentimessä on käytetty suomen kielen taivutuksen toteuttamiseen. Taivutusluokittelun käyttäminen jäsentimen lähtökohtana onnistui toteuttaa hyvin, ja siltä pohjalta muodostetun järjestelmä korpustestien perusteella taivutti oikein ne sanat mitkä sen tulikin osata, eli sanalistan luokitellut sanat. Systemaattisia puutteita sa-

nalistassa oli yhdyssanamerkintöjen kanssa, sekä sanaluokkajaossa. Luokittelussa käytetyt sanaluokat olivat vain verbejä ja nomineja, joten esimerkiksi adjektiivitaivutuksen toteutus on lopullisessa järjestelmässä vajaa.

Muita nykysuomen sanalistan taivutusluokitteluun liittyviä ongelmia olivat satunnaiset poikkeuksellisuudet osassa sanoista. Esimerkiksi sanojen *meri* ja *veri* epäsäännöllinen partitiivin vokaalisointu tai sanan *olla* epäsäännölliset taivutusmuodot vaatisivat sanalistaankin merkintänsä.

Kokonaisuutena siis sanalistan ja sen XML-formaatin osalta morfrologisen järjestelmän kannalta hyödyllistä olisi ainakin lisätä tiedot sanaluokista. Mahdollisesti järkevää olisi myös järjestellä erikoistieto muusta poikkeustaivutuksesta, vaikka toisaalta se olisi mahdollista sisällyttää myös sanaluokitusta muokkaamalla.

Kokonaisen suomen kielen kattavan jäsentimen toteuttamiseen SFST:stä löytyvä yksittäisten replace-sääntöjen kääntömahdollisuus myös riitti ja niiden käyttö sarjallistetusti sanalistatransduktoriin kompositoiden on suorituskyvyltään riittävän nopea. Tietenkin tulevaisuuden kannalta esimerkiksi mahdollisuus säännöstöjen hallintaan voi olla hyvä asia.

Korpusten osalta tilanne jäi tutkielman kannalta melko vajaaksi, sillä vapaasti ja avoimesti hyödynnettäviä testaamiseen soveltuvia korpuksia ei ollut saatavilla. Rajoitetuin ehdoin saatavilla olevista korpuksistakin kuitenkin havaittiin, että laajahko Nykysuomen sanalistakaan ei sinänsä riitä kattamaan juoksevasta tekstistä kovin suurta osaa uniikeista sanoista, vaikka sivuutettaisiinkin puutteet adjektiivitaivutuksessa ja yhdyssanamuodoissa. Sanaston ylläpitokin jää siis selvitettäväksi kysymykseksi.

Kaikkiaan tutkielman tuloksena on kuvattu järjestelmä, joka kattaa hyvän osan suomen kielen morfologisesta jäsentimestä, ja toiminee hyvänä pohjana tulevan täydemmän suomen kielen jäsentimen kehitykselle.

Viitteet

Aho, A. V., Lam, M. S., Sethi, R., ja Ullman, J. D. (2007). *Compilers: Principles, Techniques & Tools*. Pearson Addison Wesley, 2. painos.

Beesley, K. R. ja Karttunen, L. (2004). *Finite State Morphology*.

Eisner, J. (2007). Finite state software at JHU CS. `http://www.cs.jhu.edu/~jason/405/software.html`. (verkkosivu katsottu 13.4.2007).

Eronen, R. (1994). Taivutuksen osoittaminen perussanakirjassa. *Kielikello*, 1994(4).

Eronen, R. (1997). Taivutusmuotoja ikkunassa. *Kielikello*, 1997(1).

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., ja Alho, I. (2004). *Iso suomen kielioppi*. Suomalaisen Kirjallisuuden Seura.

Hakulinen, L. (1979). *Suomen kielen rakenne ja kehitys*. Otavan korkeakoulukirjasto, 4. painos.

Jurafsky, D. ja Martin, J. H. (2000). *Speech and Language Processing*. Prentice Hall.

Kanthak, S. ja Ney, H. (2004). Fsa: an efficient and flexible c++ toolkit for finite state automata using on-demand computation. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, s. 510, Morristown, NJ, USA. Association for Computational Linguistics.

Karjalainen (1992). Suomenkielinen elektroninen aineisto, joka sisältää 2 miljoonaa sanaa. Kokoojat: Helsingin yliopiston yleisen kielitieteen laitos, Joensuun yliopisto, CSC — Tieteellinen laskenta Oy. Saatavilla CSC:ltä `http://www.csc.fi`.

Karlsson, F. (1982). *Suomen kielen äänne- ja muoto-oppi*. WSOY.

Karlsson, F. (1998). *Yleinen kielitiede*. Yliopistopaino, Helsinki.

Karlsson, F., Voutilainen, A., Heikkilä, J., ja Anttila, P. (1995). *Constraint grammar—A language independent system for parsing unrestricted text*. Mouton de Gruyter.

Karttunen, L. (1995). The replace operator. In *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, s. 16—23. Cambridge, Massachusets.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word Form Generation and Recognition*. väitöskirja, Helsingin yliopisto, Helsinki. Publications 11.

Kotimaisten Kielten Tutkimuskeskus (2006). Nykysuomen sanalista. [viitattu 20.5.2007] Saatavissa `http://kaino.kotus.fi/sanat/nykysuomi`.

Lindén, K. (2005). *Word Sense Disambiguation and Discovery*. väitöskirja, Department of General Linguistics, University of Helsinki.

Remes, H. (2004). Suomen kielen äänne- ja muoto-oppi. (Luentomoniste, Joensuun yliopisto, Suomen kielen ja kirjallisuuden laitos).

Schmid, H. (2005). A programming language for finite state transducers. In *Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing (FSMNLP 2005)*. Helsinki, Finland.

Schmid, H. (2007a). *SFST Manual*.

Schmid, H. (2007b). *SFST Tutorial*.

Seppälä, S. (2006). Leksikkojen kehittämisestä äärellistilaisille morfologisille jäsentimille. Pro gradu, Helsingin yliopisto.

Setälä, E. N. (1930). *Suomen Kielioppi*. Otava, 12. painos.

Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, G., ja Varga, D. (2005). Hunmorph: open source word analysis. In *ACL 2005 workshop on software*, s. N—M.

Voutilainen, A. (1997). Designing a finite state parsing grammar. s. 283—303.

Yli-Jyrä, A. (2005). Toward a widely usable finite-state morphology workbench for less studied languages — Part I: Desiderata. *Nordic Journal of African Studies*, 14(4):479—491.

Yli-Jyrä, A. (2007). Fsmreg — a registry of FSM technology. `http://forums.csc.fi/kitwiki/pilot/view/KitWiki/FsmReg`. (verkkosivu katsottu 4.1.2008, revisio 15).

A Äärellistilaisten sovellusten ominaisuuksien vertailu

Tässä liitteessä on taulukoitu tutkielman kirjoitushetkellä löydettyjen äärellistilaisten ohjelmistojen ominaisuuksia, joiden perusteella tutkimuksessa käytetty ohjelmisto SFST on valittu, ja jonka perusteella jatkotutkimusta on mahdollista suunnitella. Taulukossa oleva data on peräisin sovellusten kotisivuilla annetuista tiedoista ja järjestelmiä kuvaavista artikkeleista. Taulukon lähtökohtana on käytetty aineistoa Yli-Jyrä (2007).

Taulukossa lisenssit on jaettu luokkiin avoimuuden perusteella. GNU-lisenssit GPL ja LGPL on merkitty erikseen. Muut avoimet lisenssit on merkitty taulukkoon vain merkinnällä *avoin*. Lisenssit, joiden ehdoilla järjestelmät ovat käytettävissä maksutta, mutta jotka rajoittavat käyttöä on merkitty *suljetuiksi*. Suuri osa suljetuista lisensseistä rajoittaa käytön nimenomaan akateemiseen ja tutkimustarkoitukseen, nämä on erikseen merkitty *tutkimus*-merkinnällä. Joitakin sovelluksia on myös markkinoilla täysin kaupallisina, siten, etteivät ne ole mitenkään vapaasti käytettävissä. Nämä on merkitty *kaupallinen*-merkinnällä.

Replace-sääntösarakkeeseen, jota on käytetty toisena tärkeänä valintaperusteena, on merkitty tukeeko järjestelmä replace-sääntöjä millään tavoin. Esimerkiksi SFST tukee yksittäisen replace-säännön muuntamista transduktoriksi ja sen kompositointia toiseen transduktoriin, mikä järjestelmän toteuttamiseen riitti, joten tätä tarkemmin en asiaa tarkastellut.

Kohtaan painollisuus on kirjoitettu tukeeko järjestelmä lainkaan painollisia transduktoreja. Kohtaan ohjelmointikielet on merkitty ensisijainen ohjelmiston ohjelmointiin käytetty kieli, joka todennäköisestikin on se, jota voi jatkokehityksessä käyttää. Jos ohjelmistoon liittyy muihin kieliin tehtyjä rajapintoja, näitä ei ole otettu huomioon. Sekä painollisuus että ohjelmointikieli lienevät jatkotutkimuksen kannalta relevantteja.

Taulukko 4: Tarkasteltavat äärellistilaiset koodikannat

Ominaisuus→ ↓Sovellus	**Lisenssi**	**Replace-säännöt**	**Painollisuus**	**Ohjelmointikieli**
AT&T FSM & GRM library	Kaupallinen Kaupallinen	Ei On	On On	C++
lextools	Kaupallinen	Ei	—	
SFST	GPL	On	Ei	C++
ALE-RA	(L)GPL	Ei	—	C++, Prolog
ASTL	LGPL	Ei	On	C++

(jatkuu seuraavalla sivulla)

Ominaisuus→ ↓Sovellus	Lisenssi	Replace-säännöt	Painollisuus	Ohjelmointi-kieli
AX	Suljettu	—	—	—
Attias FS Tools	Tutkimus	—	—	
Carmel	Tutkimus	—	—	C++, Boost
DFKI FSM Toolkit	Ei saatavilla	—	Ei	—
Edinburgh FSA	GPL	—	—	
Fadd library	Tutkimus	—	—	
FIRE Engine	Tutkimus	—	—	C++
FIRE Lite	Tutkimus	—	—	C++
FIRE Station	Tutkimus	—	—	C++
FIRE Works	Suljettu	—	—	C++
fskit	Ei saatavilla	—	—	—
GFSM	LGPL	–	On	C
GFSMT	—	–	On	LISP
Gdansk FSA	GPL	—	—	C++
GRAIL	Tutkimus	Ei	Ei	C++
Groningen FSA	GPL	–	On	Prolog
INTEX	Tutkimus	—	—	C
Jacaranda	Ei saatavilla	—	—	Java
lintouch	GPL	Ei	On	C
MAP-3.1	Kaupallinen	—	—	CLISP
MDP	—	—	—	Matlab
MMORPH	GPL	On	Ei	C
Open FIRE	Avoin	—	—	
OpenFST	Avoin	On	On	C++
PC-KIMMO	Tutkimus	Ei	Ei	C
kgen		Ei	Ei	C
REGI	Avoin	—	—	Prolog
Potsdam FSM tools	GPL	Ei	On	C++, boost
RuleCompile	Ei saatavilla	—	—	C
SRILM	Avoin	Ei	On	C++
Skeema Parser	Ei saatavilla	—	—	—
Speech Tools Library	Avoin	Ei	On	C
Statechart	—	—	—	C++, boost
TULIP	Avoin	On	Ei	Prolog
MITFST	Avoin	Ei	–	C++
RWTH FSA	Avoin	–	On	C++
UCFSM	LGPL	Ei	On	C++

(jatkuu seuraavalla sivulla)

Ominaisuus→ ↓Sovellus	**Lisenssi**	**Replace-säännöt**	**Painollisuus**	**Ohjelmointi-kieli**
UNITEX	LGPL	—	On	Java, C, C++
Vaucanson	GPL	—	On	C++
XFST	Kaupallinen	Ei	On	—
LEXC	Kaupallinen	Ei	—	
XFST2FSA	Vapaa	Ei	—	C

B Äärellistilaisten sovellusten syntaksin ja ohjeiden vertailu

Taulukko selitetty ohjeessa Eisner (2007) eli sivulla `http://www.cs.jhu.edu/~jason/405/software.html` — SFST-sarake minun, suomennokset minun. Taulukon muodon kuvaus löytyy myös viitatulta sivulta, mutta lyhyesti siinä on kerrottu jokaisen toteutuksen saatavilla olevista ohjeista ja syötteen syntaksista.

Taulukko 5: Äärellistilaisten menetelmien toteutusten vertailu

Toteutus → **↓ Asia**	**FSA Utilities**	**Xerox FST**	**AT&T FSM + lextools**	**SFST**
interactive startup	fsa -tk	xfst	—	—
manual	Holland	book	man fsmintro man lextools	 article Schmid (2007a)
complete regexp guide	Holland	sects 2.3-2.4 (p. 42ff), XRCE	 man -s5 lextools	
automaton format guide	Holland		man -s5 fsm, man -s3 fsmclass, man -s5 lextools	
interface guide	chaps 11, 12, 13, Prolog manual, Prolog-Emacs interface	overview, ref chap 3 (p. 73), help, apropos	—	—
parentheses	(E)	$[E]$	(E)	(E)
comments	% comment	# comment	# comment	% comment
atomic expressions	a, $a : b$, $a :: 3$, $a : b : 3$	a, $a : b$	a, $a : b$, $< 3 >$ ($< 3 >= \epsilon$ with weight 3)	a, $a : b$
literal symbol	$'*'$ (or escape(*))	” $*$ ” (or %*)	$\backslash*$	$\backslash*$
symbol escape codes	as in Prolog	as in C	—	$\backslash$
long symbol names	foo bar	foobar (greedy)	[foo][bar]	
complex symbol	predicates		$[nounnum = plgender = fem]$	$=pl$
symbol class	$a..z$ (or predicates)		superclass defn, in .sym file (lexmakelab compiles to .scl file for -S option)	#class#
any symbol	? (surround with spaces)	?	$[< sigma >]$ (if defined as superclass)	.
symbol complement $(? - E)$	$`E$	$\backslash E$	$\hat{}[E]$	
edge of string		$.\#.$	$[< bos >], [< eos >]$	$\# :<>$
concat	$[E1, E2, E3]$	$E1E2$	$E1E2$ (fsmconcat)	$E1E2$
character concat		$\{abcd\}$	$abcd$	$abcd$
union	$E1, E2, E3$	$E1\|E2$	$E1\|E2$ (fsmunion)	$E1\|E2$
empty string (ϵ)	[]	0 []	0 $[< epsilon >]$ (if defined)	<>
empty language	{}	$\backslash?$		
optionality	E	(E)	$E?$	$E?$
Kleene closure	$E*$	$E*$	$E*$ (fsmclosure)	$E*$
Kleene plus	$E+$	$E+$	$E+$	$E+$

Taulukko 5: Äärellistilaisten menetelmien toteutusten vertailu

Toteutus → ↓ **Asia**	**FSA Utilities**	**Xerox FST**	**AT&T FSM + lextools**	**SFST**
repetition		E^n, $E^<n$, $E^>n$, $E\{n,m\}$	E^n	
contains	$\$E$	$\$E$		
reverse	reverse(E)	$E.r$	fsmreverse	
intersect	$E1 \& E2$	$E1 \& E2$	& (fsmintersect)	$E1 \& E2$
difference	$E1 - E2$	$E1 - E2$	$E1 - E2$ (fsmdifference)	$E1 - E2$
complement ($? * - E$)	E	E	$!E$ (fsmcomplement)	$!E$
cross-product	$E1 x E2$, $E1 : E2$ (high-precedence)	$E1.x.E2$, also $abcd : aceg$	$E1 : E2$	
same-length cross-product	$E1 xx E2$			
project	domain(E), range(E)	$E.u$, $E.l$	fsmproject	$\hat{}$, _
epsilon-remove	efree(E), et al.		fsmrmepsilon	
determinize	$E!$ {t,w,wt}_determinize(E)		fsmdeterminize	
minimize	$E\#$ {t,w,wt}_minimize(E)		fsmminimize	
compose	$E1 o E2$	$E1.o.E2$	$E1@E2$ (fsmcompose)	$E1 \| \| E2$
invert	invert(E)	$E.i$	fsminvert	$\hat{}$_
ignore (insert freely)		A/B, $A./.B$ (blocked at edges)	$A << B$	
restriction		$[A => L1\_R1, L2\_R2]$	$LA => BR$	
replacement	Relevant macros in ~florian/software/lib/fsa	$A -> B$ (exhaustive nondeterm.), $A(->)B$ (optional) $A@-> B$ (LR longest) $A@> B$ (LR shortest) $A -> @B$ (RL longest) $A > @B$ (RL shortest) Reverse arrow swaps lower, upper A may have form $[.E.]$ (blocks repeat matches of epsilon in same place) B may have form $L...R$ (inserts L, R around A) Contextual restrictions: $\|\|L\_R$ (upper upper) $//L\_R$ (lower upper) $\backslash\backslash L\_R$ (upper lower) $\backslash/L\_R$ (lower lower) Parallel replacements joined by , or by ,, if restricted		$L\hat{} -> R$ $L/-> R$ $L\backslash -> R$ $L\_ -> R$