Joel Sang, Ülle Viks,
Eestin Tiedeakatemian Kielen ja Kirjallisuuden Instituutti

Tietokonella laadittu riimisanakirja

Tiivistelmä: Esitelmässä käsitellään tietokoneen käyttömahdollisuuksia viron kielen riimisanakirjan (RS) laatimisessa ja siihen liittyviä ongelmia. RS lähtee sanamuotojen sanakirjasta, joka sisältää 1,2 miljoonaa muotoa. Näistä osoittautui loppusoinnullisiksi noin 600.000. Sanakirjan volyymia supistettiin kahdella tapaa: 1) riimileikkeen siirtäminen; 2) isojen riimisarjojen redusointi. Tällä tavoin RS supistui 326.000 yksikköön. Pidemmälle automaattisilla menetelmillä ei pääse. Tietääksemme RS on ensimmäinen ja tähän asti ainoa kielen koko riimipotentiaalin esittävä sanakirja.

Esitelmämme tehtävä ei ole määritellä riimiä eikä esittää täydellistä riimien taksonomiaa. Rajoitumme seuraavassa esittämään katsauksen ainoastaan yhdestä konkreettisesta projektista – Viron kielen riimisanakirjasta, joka on edennyt viimeistelyvaiheeseen Eestin TA:n kielen ja kirjallisuuden instituutissa.

On kaksi mahdollisuutta laatia riimisanakirja. Ensimmäinen menetelmä on ottaa lähtöaineistoksi tietty määrä runotekstejä, poimia niistä kaikki riimit ja muotoilla niiden perusteella sana-artikkelit. Näin saadaan kuva todellisesta uusuksesta. Toinen mahdollisuus on lähteä jostakin olemassa olevasta sanakirjasta ja etsiä siitä kaikki loppusoinnulliset sanamuodot. Näin saadaan kuva kaikista mahdollisista riimeistä. Tässä käsitellään juuri sitä menetelmää ja siihen liittyviä ongelmia.

Jos kielessä on niukalti taivutusmuotoja, riimisanakirjan voi korvata tavallisella käänteissanakirjalla – ainoa lisäehto on se, että sanojen pitäisi olla järjestetty, ei ortografian, vaan ääntämyksen mukaan. Viron kielen käänteissanakirja ei sovellu riimisanakirjan tarpeisiin, koska sanoilla on monia taivutusmuotoja, jotka eivät tule esiin tavallisissa sanakirjoissa. Sitä vastoin sopii riimisanakirjan lähtöaineistoksi Kielen ja kirjallisuuden instituutissa vuonna 1983 valmistunut sanamuotojen sanakirja.

Sanamuotojen sanakirja (1,2 milj. muotoa) on saatu morfologisesta perussanakirjasta (36.000 sanaa) morfologisen synteesiohjelman avulla ja se sisältää kaikkien sanojen kaikki muodot (myös ne, joilla ei ole käyttöä luonnollisessa kommunikaatiossa, esimerkiksi monet *i*-monikkomuodot). Jokaiselle sanamuodolle on lisätty sen muotokoodi ja alkumuoto. Transkriptio vastaa yleensä oikeinkirjoitusta, vain morfologisesti olennaiset seikat on merkitty: III kestoaste ja ei-automaattinen paino.

```
    \'anda    Inf   'andma
   m\'aanda   Imp   m'aandama
   r\'aanda   SgG   r'aand
   t\'aanda   Imp   t'aandama
sarab.\anda   SgN   sara.banda
   v\abanda   Imp   vabandama
pr\opaganda   SgN   propaganda
   m\ajanda   Imp   majandama
   s\ajanda   SgG   sajas
  n\eljanda   SgG   neljas
    k\'anda   Inf   k'andma
    k\'anda   SgP   k'and
```

Riimisanakirja sisältää ainoastaan tasmällisiä loppusointuja. Epatäsmällisten, likimääräisten loppusointujen formalisoitua kuvausta on mahdoton esittää – missä olisi siinä tapauksessa riimin raja, kuinka suuria äänteellisiä poikkeavuuksia riimin käsite sallii?

Pääpainollisten riimien (*vaba : raba : naba...; sööksin : lööksin*) ohessa käsitellään täysiarvoisina riimeinä myös äänteellisiä vastaavuuksia pääpainollisen ja sivupainollisen tavun välillä (*kirvest\ega : sega : ega...*). Pois jäävät sivupainollisten tavujen keskinäiset vastaavuudet, jos kyseessäoleva osa ei ole loppusoinnussa pääpainollisen sanamuodon kanssa (*hirmut\atakse : kallist\atakse*).

Vaikka viron kielen ortografia on suhteellisen lähellä ääntämistä, esiintyy silti tapauksia, joissa samoin ääntyvät sanat kirjoitetaan eri tavalla ja tietokone ei osa pitää niitä riimeinä, esimerkiksi *m'aia* (SgG *maias*) ja *m'ajja* (SgIll *maja*). Jotta vältyttäisiin tältä on ennen riimien etsintää suoritettu automaattiset ortografiset muunnokset, joiden tuloksena sellaisten muotojen kirjoitusasu yhdenmukaistuu. Muunnossääntöjä on neljä:

```
R1:   ij\üj\jj → i    k'äija → k'äia
                      m'üüja → m'üia
                      m'ajja → m'aia

R2:   üüV  → üiV      hüüu → hüiu
                      m'üüa → m'üia

R3:   'VC →  'VCC     m'is → m'iss
                      ogal'ik → ogal'ikk

R4:   V\'V → 'VV      ju → j'uu
```

Toinen merkittävä ongelma on sivupaino. Koska lähtöaineistossa sivupainoa ei merkitä, niin tietokone ei pysty sitä löytämään, esimerkiksi sanassa *parem\ale (: v\ale)*. Siksi oli pakko luokitella kaikki taipuvat sanat sivupainotyyppeihin sen mukaisesti, missä sivupaino jossakin taivutusmuodossa sijaitsee (yli 100 tyyppiä).

Viron kielen painojärjestelmä on nykyisin epävakaassa siirtymävaiheessa. Tähänastinen systeemi on hajoamassa tai jo hajonnut. Kolmitavuiset sanamuodot *praeguse, raskuse, kindluse* tulevat ääntymään yhtenä painoryhmänä, josta sivupaino puuttuu. Nelitavuiset muodot *praegusele, kindlusele* ei jäsenny enää *prae-gusele, kind-lusele,* vaan *praegu-sele, kindlu-sele.* Sivupainojärjestelmä tulee yhä enemmän seuraamaan tavujen laskemisen periaatetta eikä ota huomioon kolmannen kestoasteen suurempaa painoa. Yhä enemmän astuu voimaan universaali kaksitavuisten painoryhmien tendenssi.

Mutta löytyy myös poikkeuksia ja rinnakkaismuotoja. Sanamuodon painojäsennys riippuu paljon jälkitavun rakenteesta. Näin on monikon illatiivista sanasta *juhendaja* kaksi painomallia (*juhen-daja-tesse = juhenda-jatesse*), monikon partitiivista ainoastaan yksi (*juhenda-jaid*). Sivupainon voi siirtää paikaltaan emfaattinen kolmas kestoaste: *rabelemata = rabele-mata* (= *'atta*).

Sivupainotyyppien sanaluettelot jäävät liitteeseen, riimisanakirjassa on esitetty vain sivupainolliset sananosat (noin 1800), joihin on liitetty muodon koodi ja asianomaiseen luetteloon viittava tyyppinumero.

```
"ale SgAll 2, 7-9, 17, 1-24
     PlP 60
"alile SgAll 60
"alidele PlAll 60
```

Vaikka rajoituimme ainoastaan täsmällisiin riimeihin, riimien määrä osoittautui silti hyvin suureksi – suunnilleen 600.000 (se on puolet lähtömuodoista). Supistaaksemme volyymia käytimme kahta tapaa:

a) Riimileikkeen siirtäminen. Lähtöainekseen otettiin mukaan produktiiviset pääpainolliset liitteet paradigmojensa kanssa, vastaavaliitteisissä sanamuodoissa siirrettiin riimileike jälkitavusta edemmäksi. Näin sisällytettiin riimisanakirjaan sananloppu *'aat* ja sen muodot. Kaikissa muissa taivutusmuodoissa paitsi yksikön nominatiivissa siirrettiin riimileike jälkitavusta edelliseen leikepisteeseen, siis ei *separ\aadist* vaan *s\eparaadist*. Pitempi riimipätkä tuo mukaan vähemmän riimejä: siitä sanaryhmästä (noin 140 sanaa) jää jäljelle ainoastaan yksi loppusoinnullinen pari (*s\eparaadist : pr\eparaadist*). Kaikki muut *aadist*-loppuiset muodot jätetään sanakirjasta pois. Niitä edustaa rivi *:aadist* SgEl. Osoitetun SgN perusteella sanakirjan käyttäjä voi itse muodostaa tarpeelliset muodot, sillä lähtömuodossa jäi riimileike paikalleen ja kaikki *'aat*-loppuiset sanat ovat riimeinä sanakirjassa mukana. Vastaavia pääpainollisia sananloppuja on mukana 148 (edellämainituu lisäksi *'aafia, 'ism, 'iit, 'aalne, 'eeniline, 'eerima* jne). Kuvattu menetelmä toi säästöä noin 100.000 yksikköä, volyymi supistui 520.000:een.

b) Toinen volyymin supistamiskeino on riimisarjojen redusointi. Pitemmistä riimisarjoista säilytettiin ainoastaan yksi edustaja, jos neljä ehtoa oli täytetty:

(1) kaikki sarjan jäsenet ovat samassa muodossa;

(2) se muoto ei ole lähtömuoto (SgN, Sup);

(3) jäsenten lähtömuodot ovat keskenään loppusoinnussa;

(4) sarjassa on vähintään kolme jäsentä.

Jäljelle jäänyt yksikkö varustettiin viitemerkillä, jotta sanakirjan käyttäjä osaisi lähtömuotojen perusteella itse johtaa sarjan puuttuvat jäsenet.

```
 kingi  SgG  k'ink
  ingi  SgG  l'ink
plingi  SgG  pl'ink
 mingi  SgG  m'ink
 pingi  SgG  p'ink
pringi  SgG  pr'ink
 singi  SgG  s'ink
tsingi  SgG  ts'ink          kingi ↑ SgG  k'ink

 ringi  SgG  r'ing           ringi  SgG  r'ing
svingi  SgG  sv'ing          svingi  SgG  sv'ing
```

```
 kingi  Imp  k'inkima
plingi  Imp  pl'inkima
 mingi  Imp  m'inkima
tsingi  Imp  ts'inkima        kingi ↑ Imp  k'inkima

 ringi  Imp  r'ingima         ringi  Imp  r'ingima
 tingi  Imp  t'ingima         tingi  Imp  t'ingima
```

Redusoinnin tuloksena jäi jäljelle noin 326.000 sanamuotoa (miltei 175.000 riimiä). Riimisanakirja sisältää vieläkin turhaa painolastia, mutta emme ole löytäneet automaattisia menetelmiä, joilla siitä päästäisiin eroon. Ilmeisesti tuleva työ pitää suorittaa käsin.

Ote Viron kielen riimisanakirjasta

```
          h\ale  00 hale -da
          k\ale  00 kale -da
          s\ale  00 sale -da
          p\ale  00 pale
            ale  0002 ale
          h\ale  0002 hale
          k\ale  0002 kale
          m\ale  0002 male
          p\ale  0002 pale
          v\ale  0002 vale
           "ale  07 "2, 7-9, 21-24
           "ale  16 "60
         :'aale  06 :'aa
       m\'aale↑ 06 m'aa
      •f\'aale  06•fa
      •l\'aale  06•la
  sini+r\'aale  06 sini+r'aag -a
        :'aale  16 :'aal -i
       k\'aale↑ 16 k'aal -i
       k\'aale  16 k'aal -u
       v\'aale  16 v'aal -u
 s,iia+m\'aale
s,inna+m\'aale
   sk\and'aale↑ 16 skand'aal -i
  f\arüng'aale  16 farüng'aal -i
  l\arüng'aale  16 larüng'aal -i
    tš\ek'aale  06 tšek'aa
     l\ek'aale  16 lek'aal -i
     l\ok'aale↑ 16 lok'aal -i
   b\ienn'aale  16 bienn'aal -i
  tr\ienn'aale  16 trienn'aal -i
     f\op'aale  16 fop'aa
       op'aale  16 op'aal -i
     k\op'aale  16 kop'aal -i
       or'aale  16 or'aal -i
      d\u'aale  16 du'aal -i
      v\u'aale  16 vu'aal -i
```

HAKUSANAT: riimisanakirja, riimisarja, kielen riimipotentiaali, automaattinen morfologinen analyysi, kielen painojärjestelmä