Nykysuomen sanakirjan digitaalinen editio

Niko Partanen[1][0000−0001−8584−3880] and Lotta Jalava[2][0000−0002−6400−7055]

[1] Helsingin yliopisto, Suomalais-ugrilainen ja pohjoismainen osasto
niko.partanen@helsinki.fi
[2] Kotimaisten kielten keskus
lotta.jalava@kotus.fi

Tiivistelmä Artikkeli kuvaa Nykysuomen sanakirjan näköisjulkaisun luontia ja siihen liittyviä työvaiheita. Samalla kuvataan tunnistetut rivikohtaiset tekstit ja tyylit sisältävä latauspaketti. Yhdessä ne mahdollistavat erilaisten sähköisten versioiden ja tutkimusaineistojen luomisen tulevaisuudessa, mutta ovat nykyisellään vain yksi askel tässä työssä. Tutkimus muodostaa esimerkin sanakirja-aineiston modernista tekstintunnistamisesta ja arvioi tuloksia kriittisesti, mahdollistaen samojen käytäntöjen soveltamisen muihin vastaaviin materiaaleihin. Kuvatut oikoluetut aineistot ja tekstintunnistusmallit tullaan julkaisemaan sanakirjan näköisjulkaisun rinnalla.

Avainsanat: Suomen kieli · leksikografia · OCR · saavutettavuus

1 Johdanto

Laajat ajantasaiset sanakirjat ovat osa nykyaikaisen yhteiskunnan välttämätöntä perustaa. Tarve suomen sanaston kuvaamiselle on huomattu jo varhain, ja vuosina 1880–1950 koostettiin noin 4,5 miljoonan sanalipun kokoelma Nykysuomen sanakirjan toimittamista varten. Eduskunnan vuonna 1927 tekemästä aloitteesta näiden aineistojen pohjalta julkaistiin vuosina 1951–1961 kuusiosainen Nykysuomen sanakirja, jossa on yli 200 000 hakusanaa [1,2,3,4,5,6]. Tämän jälkeen sanastotyö on jatkunut, ja 1950-luvun sanasto on jo monin osin vanhentunut. Tällä hetkellä Kielitoimiston sanakirja [7] on 100 000 hakusanallaan modernia suomen kieltä parhaiten kuvaava jatkuvasti päivittyvä lähde. Yhdessä niiden voi nähdä muodostavan laajan ja kattavan kuvauksen 1900-luvun suomen kirjakielestä. Tätä täydentävät Suomen murteiden sanakirja [9] sekä Vanhan kirjasuomen sanakirja [8], jotka molemmat ovat nykyään sähköisiä julkaisuja. Näiden kahden sanakirjan XML-muotoiset latauspaketit ovat saatavilla avoimesti CC-BY-lisensoituina.

Nykysuomen sanakirja ei ole ollut aiemmin saatavilla sähköisessä muodossa, mutta maaliskuussa 2021 Kotimaisten kielten keskus julkaisi siitä digitaalisen näköisjulkaisun Kotimaisten kielten keskuksen verkkojulkaisuja -sarjassa.[3] Saavutettavuusdirektiivin vaatimukset täyttävät PDF-tiedostot ovat nyt kaikkien käytettävissä, mutta näköisjulkaisua itseään ei ole lisensoitu täysin avoimesti.

[3] https://www.kotus.fi/nykysuomensanakirja

Sana-artikkeleja sisältävät digitoidut kuvatiedostot ja automaattisesti tunnistetut tekstit ja tyylit sisältävät tiedostot tullaan kuitenkin julkaisemaan erikseen avoimesti lisensoituna latauspakettina. Tässä artikkelissa kuvataan tämän julkaistun ja julkaistavan aineiston tekstintunnistusprosessi. Oikoluettu osuus aineistosta ja käytetyt tekstintunnistusmallit muodostavat osan latauspakettia, ja mahdollistavat niiden parantamisen ja tarkemman rakenteistamisen, joka ei itsessään kuulunut nyt tehdyn työn tavoitteisiin.

2 Aiempi tutkimus

Työ on jatkoa aiemmille julkaisuille yksittäisten kielellisesti poikkeuksellisesti haastavien aineistojen käsittelystä nykyaikaisin menetelmin. Erityisenä alan pioneerina voidaan pitää Jack Rueteria, jonka aiemmat tutkimukset ovat jo varhain tuoneet uutta tietoa erityisesti morfologisten analysaattoreiden käytöstä tekstintunnistuksessa [28], minkä lisäksi hän on julkaissut ja alkanut avata kohta vuosikymmenen tällaiselle työlle keskeistä aineistoa eri kielillä (katso mm. [24,25,26]). Viime vuosina vastaavia oikoluettuja aineistoja on julkaistu laajemminkin [18,21], ja osasta on alettu muodostaa korpuksia [17] sekä puupankkeja [19]. Tekstintunnistuksesta alkavalla työllä voi siis osoittaa olevan kauaskantoisia vaikutuksia, ja tällaiset materiaalit otetaan nopeasti uuden tutkimuksen raaka-aineiksi. Aiempia tutkimuksia edustaa myös Siperian alkuperäiskielten tekstintunnistuksen eteen tehty työ [16] sekä yhtenäiselle pohjoiselle aakkostolle laaditut tekstintunnistusmallit [20].

Työn tavoitteisiin liittyvät läheisesti myös jo tunnistetun tekstin korjauksen menetelmät ja niiden tutkimus [13,12]. Tähän Helsingin yliopistossa tehdyn työn jatkumoon sisältyy runsaasti erityisesti historiallisiin teksteihin keskittyvää tutkimusta [10,11]. Aiempi tutkimus ei kuitenkaan ole keskittynyt nimenomaisesti sanakirjojen tekstintunnistamiseen, johon liittyy hyvin erityisiä haasteita.

Yksi sanakirjojen erityispiirteistä on niiden sivulta toiselle erittäin yhdenmukaisena toistuva rakenne. Sisältö koostuu sana-artikkeleista, joissa toistuvat tietyt rakenteelliset elementit hakusanakohtaisesti laajuudeltaan vaihdellen. Sana-artikkelit on myös järjestetty aakkosittain. Tästä johtuen tietyllä tyylillä ja alkukirjaimella alkavia rivejä ei ole kuin tietyssä osassa teosta. Samanaikaisesti aineiston jakaminen riveihin ja palstoihin kuvaa sisältöä hyvin puutteellisesti, ja hyvä lähtökohta työlle olisikin esimerkiksi sana-artikkelien erottaminen jo sivun tekstialueita tunnistettaessa. Haettavan näköisjulkaisun kannalta tämä ei ollut keskeistä, joten tässä kuvatussa työssä työskenneltiin pääasiassa rivitasolla.

3 Digitoidun aineiston käsittely

Digitoidut kuvatiedostot tunnistettiin aluksi Tesseract-ohjelmalla [29] (versio 4.1.1.) suomenkielistä tekstintunnistusmallia käyttäen. Saavutettu laatu ei ollut kovin korkea, mutta Tesseractin tekstialueiden tunnistus oli erittäin tarkka, erityisesti rivikohtaisten alueiden osalta. Valitettavasti Tesseract pilkkoi palstat

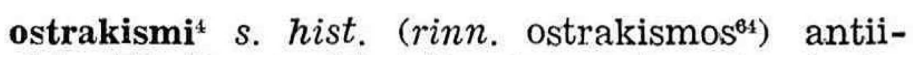
In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 177–186.

pienempiin alueisiin, jotka eivät vastanneet yksittäisiä sana-artikkeleita. Käytännössä palstat itsessään olivat yhdenmukaisesti erillä toisistaan, yksittäisiä poikkeuksia lukuun ottamatta. Nämä poikkeukset korjattiin käsin teosten lopputarkistamisen yhteydessä.

Tesseractin tunnistamat rivit muutettiin Page XML -formaattiin [22], joka vietiin Transkribus-ohjelmaan [14]. Transkribus tarjoaa erinomaisen ympäristön oikoluvulle ja eri polygonien korjaamiselle käsin, joten siinä oli vaivatonta tarkistaa yksittäisiä sivuja jokaisesta teoksesta. Myös lopullisen PDF-version ensimmäinen versio luotiin Transkribuksesta käsin.

Oikoluettujen rivien avulla tekstintunnistusta alettiin mallintaa Calamari-ohjelmaa käyttäen [30]. Tätä toistettiin siihen asti, että saavutettu laatu alkoi olla korkea ja tasainen. Viidenkymmenen oikoluetun sivun jälkeen siirryttiin vaiheeseen, jossa yksittäisten sivujen sijasta tunnistettiin koko aineisto, ja siitä valittiin jokaisesta osasta huonoiten tunnistuvia rivejä perustuen Calamari-ohjelman arvioon merkkikohtaisesta tarkkuudesta. Rajana käytettiin sitä, että jokaisen rivin täytyi sisältää ainakin yksi tai useampi alle 50% todennäköisyydellä oikein tunnistettu kirjain.

Näin menetellen pystyttiin keskittämään oikoluvun resursseja juuri näihin harvinaisiin merkkeihin, joita malli ei ollut vielä oppinut vaadittavalla tarkkuudella. Vastaavaa asteittain hankalampiin kohteisiin etenevää oikolukutyyliä olisi voinut jatkaa pidemmällekin, mutta nyt tämä lopetettiin vaiheessa, jossa tunnistamatta jääneet kirjaimet eivät pääsääntöisesti olleet harvinaisia erikoismerkkejä. Ensimmäisessä vaiheessa tämä koski erityisesti harvinaisia taivutustyyppejä, joiden yläindeksiin merkittyjä numeroita ei vielä ollut kertynyt tarpeeksi. Yläindeksin runsas käyttö on yksi kyseisen sanakirjan erityispiirteistä. Kuva 1 havainnollistaa yläindeksejä 4 ja 64.

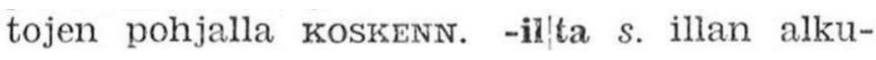


Kuva 1. Esimerkki sanojen **ostrakismi** ja **ostrakismos** taivutusnumeroista

On hyvin mahdollista, että kaikkein harvinaisimmat taivutustyypit tuottavat yhä ongelmia. Jos tekstintunnistusta haluttaisiin parantaa vielä lisää, voisi olla mahdollista tehdä taivutustyyppikohtainen oikoluku, jossa jokaisesta taivutustyypistä valittaisiin esimerkiksi sata riviä. Nyt on kuitenkin saavutettu laatu, jossa suurikaan uusien rivien oikoluku ei välittömästi paranna tekstintunnistuksen tasoa. Parannuksia voidaan kuitenkin tehdä yksittäisten merkkien osalta. Iteratiivinen oikolukuprosessi, joka etenee kokonaisista sivuista osin satunnaisesti valittuihin riveihin, ja lopulta perustuu tietoon yksittäisten merkkien ongelmista, vaikuttaa hyvin toimivalta menetelmältä laajojen teosten tekstintunnistusmallien luomiseen. Nykysuomen sanakirjassa on yli 600 000 riviä, joten oikoluvun suunnitelmallisuus oli tärkeää. Kaikkiaan oikoluettu aineisto on laajuudeltaan 8301 riviä. Tällä saavutettu tekstintunnistuksen virheprosentti on 0.36%, joten PDF-muotoisen näköisjulkaisun tarkkuus on samassa luokassa.

Yläindeksin lisäksi kirjassa käytettyjä tyylikeinoja ovat lihavointi, kursiivi ja kapiteelikirjaimet. Nämä ovat äärimmäisen tärkeitä tietoja tekstin rakenteen kannalta. Tyylien tunnistamisessa hyödynnettiin keinoa, jota on sovellettu hiljattain Daniel Sanderin *Wörterbuch der Deutschen Sprache* -teoksen digitoinnissa [23]. Näin saavutettu TEI-representaatio [23, 35] vaikuttaisi soveltuvan erittäin hyvin myös Nykysuomen sanakirjan rakenteistamiseen, ottaen huomioon, että vastaavia rakenteita on jo hyödynnetty suomalais-ugrilaisten kielten sanastotyössä [27]. Erityisen tärkeä metodologinen oivallus tässä aiemmassa työssä on ollut sanakohtaisen äänestysalgoritmin käyttö, jolla jokaisen sanan kirjasintyyli on voitu ennustaa hyvin tarkasti. Käytännössä tämän soveltaminen Nykysuomen sanakirjaan on melko suoraviivaista, sillä erolla, että osa tyyleistä voi tässä tekstissä vaihtua kesken sanan. Koska näin tapahtuu vain yläindeksin kohdalla, joka alkaa sanan lopussa ja liittyy yleensä numeromerkkien alkamiseen, voi tämän kuitenkin mallintaa ikään kuin tässä olisi myös sanaraja.

Havainnollistamme seuraavalla esimerkillä kuinka tyylitunnistetusta tekstistä voidaan päästä rakenteistetumpaan versioon. Kuvan 2 rivi sisältää kolmea eri muotoilua.



Kuva 2. Esimerkki tyylejä sisältävästä rivistä

Rivin teksti on mahdollista esittää merkkijonona *tojen pohjalla koskenn. -il|ta s. illan alku-*, joka on mitä tekstintunnistusmalli palauttaa. Sanoihin liittyvät tyylit puolestaan voi kuvata merkkijonona *xxxx xxxxxxx CCCCCCCC BBBBBB II xxxx xxxx*, tai sanatasolla voimme ilmaista, että rivillä on sanoja muotoiluin *normaali*, *normaali*, *kapiteelit*, *lihavointi*, *kursiivi*, *normaali* ja *normaali*. Käytetyt neuroverkot hyväksyvät minkä tahansa suhteen pikseleiden ja käytettyjen merkkien välillä, mutta helposti muistettavat symbolit on valittu tähän selkeyden vuoksi. Konventiot dokumentoidaan myös latauspakettiin.

Kun sanakohtaiset muotoilut ovat tiedossa on jokainen sana mahdollista ympäröidä erilaisilla rakennetta kuvaavilla elementeillä.

```
tojen pohjalla <c>koskenn.</c> <b>–il|ta</b> <i>s.</i> illan alku–
```

Tästä asusta on vielä matkaa täysin rakenteistettuun lopputulokseen, mutta se tuo tämän huomattavasti lähemmäksi. Sanakirjan jokainen sivu noudattaa samaa palstajakoa, ja rivien sijainneista on helppo selvittää niiden lukujärjestys, minkä mukaan ne on nyt järjestetty myös latauspaketissa. Suurin haaste tälle on rivien täydellisen tunnistuksen varmistaminen, johtuen rivien valtavasta määrästä, ja yksittäisiä puutteita on varmasti jäljellä. Koska rivi on kuitenkin pelkkä tietty alue kuvan polygonissa, on sen koordinaatteja helppo muuttaa. Aineiston avoin lisensointi mahdollistaa tällaisten virheiden myöhemmän korjaamisen. Tyylikohtaisia tietoja ei viety sanakirjaan, sillä tavoitteena oli tekstintunnistettu PDF-tiedosto. Oikoluetut ja tunnistetut tyylitiedot ovat kuitenkin mukana la-

tauspaketissa. Tyylien suhteen saavutettu virheprosentti on 1.52%, ja suurin osa virheistä koostui tunnistamatta jääneistä merkeistä, ei siis väärin tunnistetuista tyylikohtaisista kirjaimista. Tyylimallin oikoluetun aineiston laajuus on 5062 riviä kattaen sivuja kaikista kuudesta osasta. Tyylimallin pienempi koko johtui tyyleihin tehtävän annotointityön hitaudesta – mainittakoon myös, että kaikki tyylimallin rivit ovat myös tekstintunnistusmallissa tekstisisältönsä puolesta.

4 Saavutettava näköiskappale

Nykysuomen sanakirjan PDF-muotoinen digitaalinen näköiskappale on tekstintunnistettu yllä kuvatuin menetelmin. Jokaiseen tiedostoon on tehty Adobe Acrobat Pro -ohjelmalla automaattinen tekstielementtien luokittelu, minkä jälkeen ne on tallennettu optimoituina PDF-tiedostoina. Optimointi muutti käytännössä tekstielementtien tagien tallentamistapaa, ja pienensi PDF-tiedostoja huomattavasti. Tiedostojen käyttö testatusti onnistuu myös ruudunlukulaitteilla, joten ne voidaan laskea saavutettavuudeltaan vähintään kohtuullisiksi. Tiettyjä puutteita on silti yhä löydettävissä, ja esimerkiksi sivulta toiselle navigoidessa joutuu ainakin tietyillä ruudunlukijoilla pakottamaan sivunvaihdon erikseen. Tälle ei tässä vaiheessa löydetty parempaa ratkaisua. Saavutettavuuden kannalta paras vaihtoehto tuleekin olemaan kokonaan rakenteistetun sähköisen version luominen, josta erilaiset haut ja suodatukset onnistuvat vaivatta.

Tekstintunnistuksen tarkkuus on erittäin korkea. Sitä arvioidessa on myös otettava huomioon, että oikoluettavaan aineistoon on erityisesti sisällytetty haastavia rivejä. Tämä laskee mallin kokonaistarkkuutta testatessa, mutta varmistaa realistisemman ja yhdenmukaisemman tarkkuuden koko aineistossa.

Aineiston käsittelyssä tehtiin muutama tekninen ratkaisu, jotka on huomioitava myös jatkokäytössä. Erilaiset viivat on yhdenmukaistettu yksinkertaisiksi väliviivoiksi. Oikoluetussa materiaalissa ne ovat erillään, mutta koska järjestelmä ei tunnistanut niiden eroa tyydyttävällä tavalla, ei tätä eroa huomioitu PDF-tiedostossa. On kuitenkin selvää, että myöhemmän rakenteistamisen kannalta tämä ei ole ihanteellinen ratkaisu. Lainausmerkit on tarkoituksella tuotettu yhtenä tai kahtena heittomerkkinä, mikä johtuu Calamari-ohjelman sisäisestä normalisoinnista. Käytännössä tämänkin eron voisi säilyttää eri versioissa niin haluttaessa. Yksittäisten erikoismerkkien suhteen on varmasti runsaasti perusteltuja vaihtoehtoja, eivätkä nykyiset valinnat ole välttämättä täydellisiä. Saavutettavuuden kannalta olisi syytä huomioida erityisesti se, kuinka eri merkit toimivat ruudunlukijoilla.

Muun käsittelyn jälkeen tiedostoihin on luotu myös kirjainkohtaiset kirjanmerkit PDF-tiedostojen navigointia helpottamiseksi. Tämän käytettiin pikepdf Python-kirjastoa.[4]

[4] https://github.com/pikepdf/pikepdf

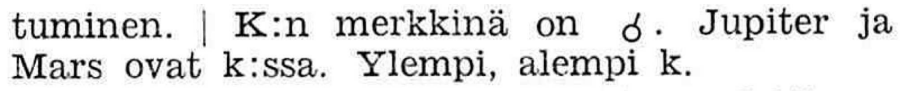
5 Jäljelle jääneet puutteet ja työvaiheet

Yksi näin luotujen PDF-tiedostojen puutteista on niiden suuri koko. Käytännössä tämä aiheutuu niihin sisällytettyjen kuvatiedostojen suuresta koosta. Paras keino pienentää tiedostoja olisikin esimerkiksi kuvien binarisointi. Sitä ei tehty tässä vaiheessa, sillä tekstintunnistuksen kannalta alkuperäinen harmaasävykuva oli paras vaihtoehto. Binarisointi voisi haitata tiettyjen voimakkaasti häivettyneiden merkkien lukua. Koska tekstiä ei ole nyt kokonaisuudessaan oikoluettu, ei tällainen mahdollisesti luettavuutta estävä toimenpide ole välttämättä toivottu. Teknisesti tekstintunnistuksen yhdistäminen esimerkiksi binarisoituihin sivuihin olisi täysin mahdollista, ja onkin yksi lisätyötä tarvitseva toimenpide.

Vaikka rivit ovat tunnistuneet yleisesti ottaen oikein, on tilanteita, joissa niihin liittyy ongelmia. Yksittäisissä kohdissa rivi on joko jakaantunut kahteen osaan tai pieni osa siitä puuttuu. Erityisesti jälkimmäiset tilanteet vaikuttavat erittäin harvinaisilta. Tämä kuitenkin vaikuttaa palstojen rakenteistamiseen, joten mahdolliset poikkeukset rivien sijainneissa on huomioitava jatkokäsittelyssä. Myös sivujen yläreunassa olevat sivunumerot ja kirjainvälin lyhenteet on usein tunnistettu samaksi tekstialueeksi. Tämä on käytetyn alueidentunnistusmallin ominaisuus, eikä sitä korjattu erikseen, mutta sen vaikutus esimerkiksi sivunumeroiden tunnistamiseen on hyvä huomioida.

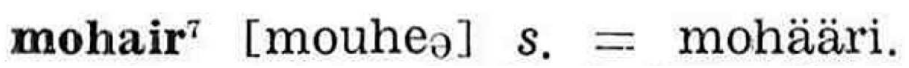
Tiettyjen merkkien harvinaisuus aineistoissa aiheuttaa niiden epävarman tunnistuksen. Niiden tunnistuminen on mahdollista, sillä niiden sisällyttämiseen malliin kiinnitettiin runsaasti huomiota. Silti varma tunnistustarkkuus vaatisi jokaisesta merkistä useamman esimerkin eri konteksteissa. Tietyistä merkeistä emme voi olla varmoja, esiintyvätkö ne kirjassa edes useampia kertoja. Näin ollen erityisesti harvinaisten suomen kirjakielelle vieraiden foneettisten merkkien tarkistus tulee vaatimaan työtä myös tulevaisuudessa. Mainittakoon, että jäljelle jääneet virheet ovat usein melko pieniä, ja esimerkiksi merkki $\bar{u}$ tunnistuu ajoittain merkkinä u tai $\ddot{u}$. Tämä on puute, muttei välttämättä kriittisesti rajoita kirjan käytettävyyttä. Yksittäisillä sivuilla on myös erilaisia tahroja, joiden osalta tekstintunnistus täytyisi tarkistaa kirjan toisesta kappaleesta.

Muita ongelmia tuottaneita merkkejä voivat olla sellaiset, jotka esiintyvät tekstistä äärimmäisen harvoin. Rajana voidaan pitää, että yksittäisin merkin täytyy esiintyä harjoitusaineistossa ainakin muutamia kertoja, jotta malli voisi edes teoriassa tunnistaa sitä. Tällaisesta erittäin harvinaisesta merkistä voimme ottaa esimerkiksi astronomisen konjunktion symbolin. Se ei välttämättä esiinny kirjasarjassa kuin kerran. Toisaalta tätä on mahdoton todistaa, ennen kuin koko teos on oikoluettu. Nyt tällaisten harvinaisten merkkien paikalle on laitettu oikea Unicode-merkki.


tuminen. | K:n merkkinä on ☌. Jupiter ja
Mars ovat k:ssa. Ylempi, alempi k.

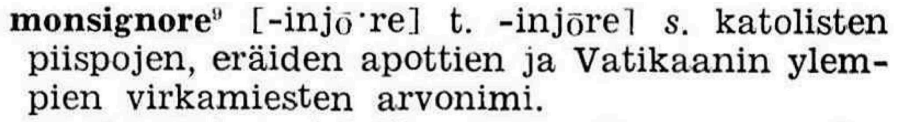
Kuva 3. Esimerkki konjunktion symbolin käytöstä

Hieman useamminkin esiintyvä harvoin esiintyvä merkki voi olla erityisen altis tunnistuksen horjunnalle, jos painojäljessä on pienintäkään kulumaa tai vaihtelua. Valitettavasti juuri harvinaisten merkkien kohdalla tällainen on yleistä. Tästä esimerkkinä voidaan ottaa sanan **mohair** artikkelin ensimmäinen rivi, jossa švaa itse asiassa oli tunnistunut oikein, mutta näemme Calamari-mallin arviosta, ettei kovin varmasti.



Kuva 4. Ensimmäinen rivi sana-artikkelista **mohair**

Esimerkkinä epäonnistuneesta tunnistuksesta voimme ottaa sanan **monsignore** artikkelin Kuvassa 5, jossa kirjain *ō* on tunnistunut kirjaimena *ö*, kuten jo yllä mainittiin yleisenä ongelmana. Näemme tässä myös poikkeuksellista kulumaa pituusmerkin kohdalla, mikä on osaltaan voinut vaikuttaa tunnistuksen lopputulokseen.



Kuva 5. Sana-artikkeli **monsignore**

Vielä laajemmalla oikoluennalla, jossa erityisesti keskityttäisiin tällaisiin ongelmallisiin riveihin, olisi mahdollista parantaa tunnistusta entisestään. Koska ääntämisohjeita sisältäviä näitä rivejä on teoksissa rajallinen määrä, voisi käsin tehtävän korjauksen ulottaa erityisesti näihin riveihin.

Yksi keskeinen jatkokäytön toimenpide koskee eri teoksissa olevien korjauslistojen sisällyttämistä itse tekstiin. Ne voisi esimerkiksi lisätä sähköiseen versioon kommentteina, ja ottaa huomioon rakenteistetussa verkkoversiossa. Näin ollen sähköinen versio ei olisi näköisjulkaisu, kuten nyt julkaistut PDF-versiot, vaan toisi yhteen julkaistut aineistot alkuperäisten tekijöiden tarkoittamien korjausten kera.

Sarjan viimeisessä niteessä [6, 777-810] on myös laaja Täydennyksiä-osio. Voi ajatella, että rakenteistetussa versiossa nämä täydennykset voitaisiin viedä oikeille paikoilleen artikkelien sekaan, maininnalla niiden täydennyksellisestä roolista ja sijainnista kuudennessa niteessä. Nykysuomen sanakirjahan on valmis työ, johon ei voi tehdä muutoksia. Yllä kuvattujen toimien avulla voisi kuitenkin muodostaa verkossa olevan rakenteistetun Nykysuomen sanakirjan digitaalinen edition, johon nyt tehdyn työn toivotaan johtavan.

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 177–186.

6 Aineiston jatkokäytön mahdollisuuksista

Nykysuomen sanakirjan sähköisillä versioilla tulee eittämättä olemaan merkittävä rooli suomen kielen tutkimuksessa 2000-luvulla, kuten sen painetulla versiolla on ollut edeltävällä vuosisadalla. Aineisto tarjoaa rajattomasti mahdollisuuksia erityisesti sanaston muuttumisen tutkimukseen. Samanaikaisesti jatkuvasti paremmin saataville tulevat korpukset esimerkiksi suomenkielisistä sanomalehdistä [15] ja kirjallisuudesta 1900-luvulla mahdollistavat eri aikoina tehdyn sanakirjatyön tarkastelun myös näiden aineistojen valossa. Erityisesti sanakirjan tekovaiheessa syntyvä sanasto on voinut vakiintua eri muodoissa, mutta ehkä sanakirja on itse voinut vakiinnuttaa tiettyjen sanojen käyttöä. Tätä voisi tutkia vertailemalla kilpailevien varianttien frekvenssiä esimerkiksi vuositasolla. Myös Nykysuomen sanakirjan laaja aihealueittainen luokittelu voi mahdollistaa tarkan alakohtaisen terminologian muutosnopeuksien vertailun.

7 Yhteenveto

Suomessa 1900-luvun aikana julkaistujen sanakirjojen digitointi ja rakenteistaminen on yksi niistä tehtävistä, joille tulevien vuosikymmenten tutkijat tulevat väistämättä omistamaan runsaasti aikaansa. Tärkeimmät suomenkieliset sanakirjat alkavat olla sähköisesti saatavilla, mutta toisin on esimerkiksi suomalais-ugrilaisista kielistä tehtyjen sanakirjojen kohdalla.

Nykysuomen sanakirjan näköisjulkaisun tekstintunnistus osoitti, että erittäin korkealaatuinen jälki on mahdollista saavuttaa kohtuullisella vaivalla. Eniten ongelmia aiheutuukin rivien ja muiden alueiden tunnistuksesta, ja kehitys tällä saralla olisi erittäin toivottavaa. Sanakirja tarjoaa aineistona omanlaisia haasteitaan, muttei mitään sellaista, mistä aiheutuisi ylitsepääsemättömiä ongelmia. Näin ollen vastaaviin hankkeisiin on erittäin suositeltavaa ryhtyä myös muiden julkaistujen vastaavien aineistojen kohdalla.

On myös mainittava, että Nykysuomen sanakirjaan liittyy laaja kortistoaineisto, jota ei ole digitoitu. Myös tällaisten aineistojen jatkokäytön mahdollisuuksia on äärimmäisen tärkeä selvittää. Samanaikaisesti ainoastaan tällaisten aineistojen jatkuva kehittäminen ja avoin julkaisu varmistaa, että materiaalit hyödyttävät näiden suomen puhujia, tutkijoita ja muita käyttäjiä parhaalla mahdollisella tavalla.

Viitteet

1. Nykysuomen sanakirja. Ensimmäinen osa. A–I. Werner Söderströmin Osakeyhtiö (1951), päätoimittaja: Matti Sadeniemi. Toimitussihteeri: fil. maist. Jouko Vesikansa. Toimittajat: fil. maist. Simo Hämäläinen, fil. maist. Arvo Keinonen, fil. maist. Ritva Peltonen fil, maist. Taito Piiroinen, fil. tri Paavo Siro, fil. maist. Hannes Teppo.

2. Nykysuomen sanakirja. Toinen osa. J–K. Werner Söderströmin Osakeyhtiö (1953), päätoimittaja: Matti Sadeniemi. Toimitussihteeri: fil. maist. Jouko Vesikansa. Toimittajat: fil. maist. Simo Hämäläinen, fil. maist. Arvo Keinonen, fil. maist. Taito Piiroinen, fil. tri Paavo Siro, fil. maist. Hannes Teppo, fil. maist. Jorma Vuoriniemi.
3. Nykysuomen sanakirja. Kolmas osa. L–N. Werner Söderströmin Osakeyhtiö (1954), päätoimittaja: fil. tri Matti Sadeniemi. Toimitussihteeri: fil. maist. Jouko Vesikansa. Toimittajat: fil. maist. Simo Hämäläinen, fil. maist. Arvo Keinonen, fil. maist. Taito Piiroinen, fil. maist. Hannes Teppo †. fil. maist. Jouko Vahe, fil. maist. Jorma Vuoriniemi.
4. Nykysuomen sanakirja. Neljäs osa. O–R. Werner Söderströmin Osakeyhtiö (1956), päätoimittaja: dos. Matti Sadeniemi. Toimitussihteeri: fil. maist. Jouko Vesikansa. Toimittajat: fil. maist. Simo Hämäläinen, fil. maist. Arvo Keinonen, fil. maist. Taito Piiroinen, fil. kand. Paavo Pulkkinen. fil. maist. Jouko Vahe, fil. maist. Jorma Vuoriniemi.
5. Nykysuomen sanakirja. Viides osa. S–Tr. Werner Söderströmin Osakeyhtiö (1959), päätoimittaja: dos. Matti Sadeniemi. Toimitussihteeri: fil. maist. Jouko Vesikansa. Toimittajat: fil. maist. Arvo Keinonen, fil. lis. Pentti Lyly, fil. maist. Osmo Nikanne. fil. kand. Liisa Nurmela, fil. maist. Taito Piiroinen.
6. Nykysuomen sanakirja. Kuudes osa. Ts–Ö. Werner Söderströmin Osakeyhtiö (1961), päätoimittaja: dos. Matti Sadeniemi. Toimitussihteeri: fil. maist. Jouko Vesikansa. Toimittajat: fil. maist. Arvo Keinonen, fil. maist. Osmo Nikanne, fil. kand. Liisa Nurmela, fil. maist. Kaarina Åstedt.
7. Kielitoimiston sanakirja. Kotimaisten kielten keskus (2004–2021), https://www.kielitoimistonsanakirja.fi, päätoimittaja: Tarja Riitta Heinonen. Toimitussihteeri: Minna Haapanen. Toimittajat: Leena Joki, Riina Klemettinen, Ilona Paajanen, Minna Pyhälahti
8. Vanhan kirjasuomen sanakirja. Kotimaisten kielten keskus (2014–2020), https://kaino.kotus.fi/vks/, päätoimittaja: Maria Lehtonen. Toimitussihteeri: Katariina Summanen. Sanakirjaosaston johtaja: Pirkko Kuutti. Toimittajat: Elina Heikkilä, Kalle Järvelä, Jarkko Kauppinen, Taru Laanti
9. Suomen murteiden sanakirja. Kotimaisten kielten keskus (2021–2020), https://kaino.kotus.fi/sms/, päätoimittaja: Heikki Hurtta ja Nina Kamppi. Sanakirjatoimittajat: Kirsti Aapala, Tarja Korhonen, Anna Ryödi, Minna Salonen, Riikka Tervonen, Eeva Tuominen
10. Drobac, S., Lindén, K.: Optical character recognition with neural networks and post-correction with finite state methods. International Journal on Document Analysis and Recognition (IJDAR) **23**(4), 279–295 (2020)
11. Drobac, S., Lindén, K., et al.: Optical font family recognition using a neural network. In: Proceedings of the Research Data And Humanities (Rdhum) 2019 Conference Data, Methods And Tools. Studia Humaniora Ouluensia (2019)
12. Duong, Q., Hämäläinen, M., Hengchen, S.: An unsupervised method for ocr post-correction and spelling normalisation for finnish. arXiv preprint arXiv:2011.03502 (2020)
13. Hämäläinen, M., Hengchen, S.: From the paft to the fiiture: a fully automatic nmt and word embeddings method for ocr post-correction. arXiv preprint arXiv:1910.05535 (2019)
14. Kahle, P., Colutto, S., Hackl, G., Mühlberger, G.: Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 4, pp. 19–24. IEEE (2017)

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 177–186.

15. Kansalliskirjasto: Kansalliskirjaston sanoma- ja aikakauslehtikokoelman suomenkielinen osakorpus, Kielipankki-versio (2011), http://urn.fi/urn:nbn:fi:lb-2016050302
16. Partanen, N.: Challenges in ocr today: Report on experiences from inel. In: Elektronnaya Pis'mennost' Narodov Rossiyskoy Federatsii: Opyt, Problemy I Perspektivy. pp. 263–273 (2017)
17. Partanen, N.: langdoc/four-battles-corpus: Parallel corpus for Erzya, Hill Mari, Permian Komi, Zyrian Komi and Udmurt (Mar 2019). https://doi.org/10.5281/zenodo.2615038
18. Partanen, N.: nikopartanen/vyl-tujod-ocr: Vyl' Tujöd newspaper Ground Truth (May 2019). https://doi.org/10.5281/zenodo.3232996
19. Partanen, N., Blokland, R., Lim, K., Poibeau, T., Rießler, M.: The first Komi-Zyrian Universal Dependencies treebanks. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pp. 126–132. Association for Computational Linguistics (2018), http://aclweb.org/anthology/W18-6015
20. Partanen, N., Rießler, M.: An ocr system for the unified northern alphabet. In: The fifth International Workshop on Computational Linguistics for Uralic Languages (2019)
21. Partanen, N., Rießler, M.: langdoc/iwclul2019: An OCR system for the Unified Northern Alphabet – data package (Dec 2018). https://doi.org/10.5281/zenodo.2506881
22. Pletschacher, S., Antonacopoulos, A.: The page (page analysis and ground-truth elements) format framework. In: 2010 20th International Conference on Pattern Recognition. pp. 257–260. IEEE (2010)
23. Reul, C., Göttel, S., Springmann, U., Wick, C., Würzner, K.M., Puppe, F.: Automatic semantic text tagging on historical lexica by combining ocr and typography classification: A case study on daniel sander's wörterbuch der deutschen sprache. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage. pp. 33–38 (2019)
24. Rueter, J.: Initial ocr word form list for scanning of ingrian, 1930s (2012-11-06) (Nov 2012)
25. Rueter, J.: rueter/shoks_kolkhoznikinj-valjgij-1932-33: Shoksha-languoid Kolkhoznikin' val'gij newspaper articles 1932–1933 (Feb 2018). https://doi.org/10.5281/zenodo.1165766
26. Rueter, J.: rueter/Erzya-grammar-Wiedemann-1865: Initial release of proofread Erzya Grammar by Wiedemann 1865 (Sep 2019). https://doi.org/10.5281/zenodo.3385183
27. Rueter, J., Hämäläinen, M., et al.: On xml-mediawiki resources, endangered languages and tei compatibility, multilingual dictionaries for endangered languages. In: AsiaLex 2019 Proceedings of the 13th Conference of the Asian Association for Lexicography. Asos Publisher (2019)
28. Silfverberg, M., Rueter, J.: Can morphological analyzers improve the quality of optical character recognition? In: Septentrio Conference Series. pp. 45–56. No. 2 (2015)
29. Smith, R.: An overview of the tesseract ocr engine. In: Ninth international conference on document analysis and recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
30. Wick, C., Reul, C., Puppe, F.: Calamari-a high-performance tensorflow-based deep learning package for optical character recognition. arXiv preprint arXiv:1807.02004 (2018)