Kimmo Kettunen, Tuula Pääkkönen & Mika Koistinen

Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910: sanatason laatu, kokoelmien käyttö ja laadun parantaminen

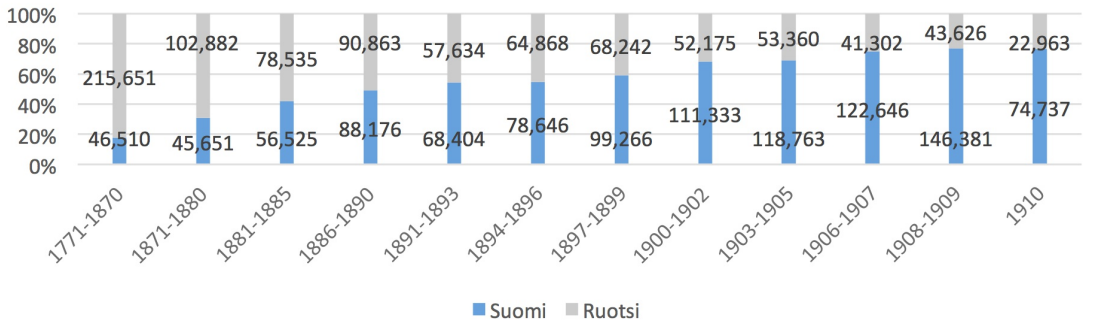
Kimmo Kettunen, Kansalliskirjasto, kimmo.kettunen@helsinki.fi;
Tuula Pääkkönen, Kansalliskirjasto, tuula.paakkonen@helsinki.fi;
Mika Koistinen, Kansalliskirjasto, mika.koistinen@helsinki.fi

Erilaista vanhaa tekstiaineistoa on digitoitu usean vuosikymmenen ajan ja materiaalia on saatavilla eri kielillä huomattavia määriä. Digitoitujen vanhojen tekstien kokonaismäärien arvioiminen on vaikeaa, mutta muutaman esimerkin mainitseminen riittää kertomaan aineistojen laajuudesta. Europeana-projekti arvioi vuonna 2012, että digitoituja vanhoja sanomalehtiä oli Euroopan tasolla olemassa noin 24 000 nimikettä ja yli 128 miljoonaa sivua (Dunning, 2012). British Libraryn 1800-luvun lehtikokoelmassa 19th Century British Library Newspapers Database on noin 3 miljoonaa sivua ja 70 nimikettä. Gutenberg-projekti ilmoittaa, että heillä on saatavilla yli 50 000 digitoitua kirjaa. Yhdysvaltalaisella Hathitrustilla on digitoituna 14 558 573 nidosta, joissa on 5 095 500 550 sivua. Kansallisarkisto ilmoittaa digitoitujen aineistojensa määräksi kirjoitushetkellä 42 244 873. Käytetyt mitat eivät välttämättä ole suoraan verrannollisia keskenään, mutta luvuista saa käsityksen aineistojen määrästä.

Kansalliskirjasto ilmoittaa verkkosivuillaan (digi.kansalliskirjasto.fi, tästä lähtien Digi) digitoitujen sivujen kokonaismääräksi kirjoitushetkellä vuoden 2016 syyskuussa 10 381 173. Luku sisältää sanoma- ja aikauslehdet sekä pienpainatteet. Koko lehtiaineistosta vuodet 1771–1910 ovat vapaasti käytettävissä, ja niissä on yhteensä noin 3 miljoonaa sivua. Sanomalehtiä on 445 nimikettä, aikakauslehtiä 3141.

Tässä artikkelissa käsitellään pääsääntöisesti Kansalliskirjaston suomenkielisen digitoidun sanoma- ja aikakauslehtikokoelman laatua sanatasolla sekä koko aineiston käyttöä yleisesti. Laatuanalyysi kattaa vuodet 1771–1910, jakson jonka aineisto on vapaasti käytettävissä kirjaston digitaalisten aineistojen verkkosivulla. Aineisto on myös saatavilla vuoteen 1874 asti Kielipankissa sekä käytettävissä kokonaisuudessaan kielentutkimuksen Korp-palvelussa. Koko aineisto vuosilta 1771–1910 on tulossa myöhemmin vuonna 2016 avoimen datan jakelupakettina saataville (Pääkkönen ja kumppanit, 2016).

Lehtiaineisto on pääasiassa suomen- ja ruotsinkielistä, mutta seassa on myös vähäisiä määriä saksaa, venäjää, ranskaa ja muita kieliä. Kuvassa 1 esitetään sanomalehtiaineiston jakautuminen suomen- ja ruotsinkieliseen materiaaliin sivuina. Aikavälit kuvassa ovat aineistosta



Kuva 1: Suomen ja ruotsin osuus sanomalehtiaineistossa sivuina

tehdyn avoimen datan jakelupaketin mukaisia (Pääkkönen ja kumppanit, 2016). Vuoteen 1890 saakka aineiston enemmistö on ruotsinkielistä, siitä eteenpäin suomeksi julkaistaan enemmän. Suomenkielisten sivujen kokonaismäärä on 1 063 648, ruotsinkielisten 892 101.

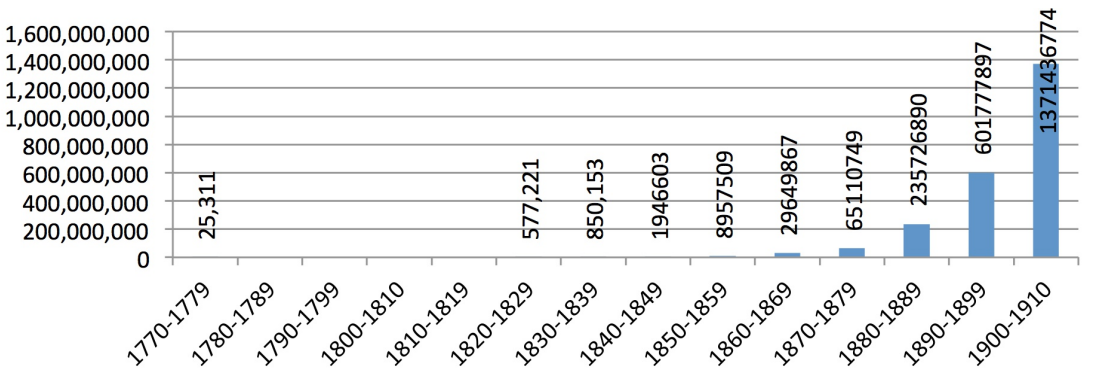
Esitämme artikkelissa laatuanalyysin suomenkielisestä aineistosta, ruotsinkieliseen emme ole toistaiseksi paneutuneet. Vastaavan analyysin voi kuitenkin tehdä artikkelissa ja lähteissä esitettyä prosessia noudattaen.

Kansalliskirjaston sanoma- ja aikakauslehtiaineiston suomenkielisessä osassa on noin 2.407 miljardia sanaa. Vuosikymmenittäin jaoteltuna aineisto jakautuu kuvan 2 mukaisesti.

Suurin osa aineistosta, 82.7 %, on julkaistu ajanjakson kahtena viimeisenä vuosikymmenenä, vuosina 1890–1910. Neljä viimeistä vuosikymmentä, vuodet 1870–1910, kattavat aineistosta 92.3 %.

Digitoinnin ongelmat

Laajat vanhojen tekstiaineistojen digitoinnit törmäävät monesti siihen, että digitoinnin tulos ei ole kovin hyvä: lopputuloksena syntyvään digitaaliseen tekstiin jää paljon virheitä. Tämä johtuu monesta tekijästä: huonosta alkuperäisten lehtien painolaadusta, kehnosta paperista, mikrofilmin laadusta, skannauksen resoluutiosta, hankalista kirjasinlajeista jne. (vrt. Holley 2008; Klijn 2008; Piotrowski, 2012). Esimerkiksi British Libraryn 19th Century Newspaper Project on arvioinut noin yhden prosentin otoksella kahden miljoonan projektissa digitoidun pai-



Kuva 2: Lehtiaineiston sanamäärät vuosikymmenittäin; vuosina 1780–1819 julkaistiin vain ruotsinkielisiä lehtiä

nosivun laadun ja saanut tulokseksi, että noin 78 % aineiston sanoista on oikein, loput virheellisiä (Tanner ja kumppanit, 2009). Tulosta ei voi pitää laadullisesti hyvänä, mutta se on realismia. Niklasin (2010) aineisto kattaa The Times of Londonin digitoidut vuosikerrat 200 vuoden ajalta vuosilta 1785–1985.Aineistossa on noin 7 miljardia sananmuotoa ja 8 miljoonaa artikkelia. Laatuarviot aineistosta on tehty käyttäen digitaalisia sanakirjoja. Tällainen laatuarvio on kattava, koska siinä arvioidaan koko aineisto, mutta lopputulos on sanan täydessä merkityksessä arvio: käytetyt digitaaliset sanakirjat vaikuttavat tulokseen monin tavoin. Ne eivät voi sisältää kaikkia sanoja koskaan, ja toisaalta vaikka sana olisi sanakirjan tunnistama, voi se olla jotain muuta, kuin alkuperäisessä tekstissä. Sanojen tunnistettavuus The Times of Londonin aineistossa vaihtelee 55:n ja 80 prosentin välillä. Kautta koko 1900-luvun sanojen tunnistettavuus pysyttelee pääasiassa 70–80 prosentissa.

Tekstin virheiden vaikutukset

Digitoidun tekstin virheet aiheuttavat monenlaisia vaikeuksia. Tekstin lukemisen ja ymmärtämisen vaikeutuminen on niistä ilmeisin. Sen lisäksi virheet hankaloittavat tekstien löytämistä hakujärjestelmistä: on mahdollista että tekstiä ei löydetä hakusanojen perusteella ollenkaan tai tekstiosuma sijoittuu tuloslistassa niin alas, että käyttäjä ei huomaa sitä (Taghva ja kumppanit, 1996; Kantor ja Voorhees, 2000; Mittendorf ja Schäuble, 2000). Tiedonhaun koeasetelmissa hakutulosten heikkenemiset ovat olleet melko dramaattisia jo 5:n ja 20 prosentin virhemäärillä (Kantor ja Voorhees, 2000; Savoy ja Naji, 2011), jotka ovat hyvin yleisiä digitoiduissa teksteissä. Toisaalta hakujärjestelmät ovat myös hyvin joustavia ja kykenevät löytämään dokumentteja roskaisestakin aineistosta (Mittendorf ja Schäuble, 2000). Suurin ongelma ovat lyhyet kyselyt ja lyhyet dokumentit, joissa sanat eivät toistu riittävän usein, jotta hakujärjestelmä löytäisi dokumentin. Myös merkkivirheiden jakauma aineistossa vaikuttaa hakuun: tasaisesti sanoihin jakaantuneet merkkivirheet aiheuttavat vähemmän ongelmia (Mittendorf ja Schäuble, 2000). Niin sanotusta sumeasta merkkijonojen täsmäytyksestä on apua virheellisen tekstin etsinnässä, mutta kaikkia tekstin virheitä sekään ei kykene selvittämään, ja hakutulokset voivat jäädä vaatimattomiksi (Järvelin ja kumppanit, 2015).

Tiedonhaun ongelmat ovat yksi esimerkki kehnon optisen luvun laadun aiheuttamista ongelmista. Erilaisille kieliteknologian ja tekstinlouhinnan sovelluksille virheet aiheuttavat myös ongelmia (Lopresti, 2009). Yhtenä esimerkkinä tästä voi mainita nimien tunnistuksen ja eristämisen aineistoista (named entity recognition, NER, Nadeau ja Sekine, 2007). Tutkimuskirjallisuudesta tiedetään, että esimerkiksi nimien tunnistaminen aineistoista vaikeutuu, jos teksti on kovin virheellistä (esimerkiksi Alex ja Burns, 2016; Packer ja kumppanit, 2011). Omat tuloksemme vahvistavat tätä käsitystä. Kettunen ja kumppanit (2016) tutkivat nimien tunnistamista lehtiaineistosta tehdyssä evaluaatiokokoelmassa. Saavutetut tulokset olivat melko vaatimattomia, parhaimmillaan vain noin 60 % nimistä tunnistettiin, ja yleensä tulokset olivat 30–50 prosentin tasolla tai vielä alempia. Pienellä korjatulla aineistolla tehty koe osoitti, että nimien löytyvyys parani, kun tekstin laatua oli saatu parannetuksi.

Aineistojen yleinen käytettävyys

Tiedonhaun tutkimuksessa roskaisen aineiston tutkimus on keskittynyt paljolti virheiden vaikutukseen hakujärjestelmien suorituskykyyn. Digitoitujen tekstien virheiden tutkijalle aiheuttamista käytännön ongelmista antavat esimerkkejä muun muassa Traub ja kumppanit (2015) sekä Hitchcock (2013). Traub ja kumppanit haastattelivat historiantutkijoita ja kysyivät heidän tutkimusongelmiaan. Sen perusteella he arvioivat, miten hyvin lehtiaineiston hakujärjestelmät kykenivät auttamaan tutkimuksessa. Joissain tutkimuskysymyksissä järjestelmät toimivat, joissain eivät. Tutkijat ovat yleisesti ottaen tietoisia optisen luvun aiheuttamista aineiston virheistä, mutta eivät itse kykene arvioimaan, miten ne vaikuttavat heidän tutkimuksensa suorittamiseen. Tässä saattaisi auttaa tarkempi tieto digitointiprosessista ja tilastotiedot tekstin virheiden määrästä.

Digitoidut lehtiaineistot ovat tärkeitä tutkimuslähteitä muun muassa historiantutkimukselle. Esimerkiksi vuonna 2016 alkaneessa Suomen Akatemian rahoittamassa hankkeessa Digitaalinen historiantutkimus ja julkisuuden muutos Suomessa 1640–1910 Kansalliskirjaston lehtiaineisto on yksi keskeisistä tutkimuksen lähteistä.

Kansallisarkiston ja Kansalliskirjaston digitoitujen aineistojen käyttökyselyssä (Hölttä, 2016) vastaukset saatiin 112 henkilöltä, joista tutkimuskokemusta oli 24 prosentilla ja yli 10 vuotta tutkimuskokemusta 13 prosentilla vastaajista. Kyselyn tuloksissa käyttäjien esiin nostama pääsyy digitaalisten aineistojen käyttämättömyyteen ei ollut aineiston laatu vaan se, että aineiston vapaa verkkokäyttö ei ole mahdollista joko siksi että aineistoa ei ole vielä digitoitu tai että siinä on käyttörajoituksia (Hölttä, 2016). Tämä on selkeästi nähtävissä myös Digin käyttäjäpalautteista, joista vuonna 2015 noin 25 % käsitteli joko aineiston ajallisen käytön takarajan pidentämistä tai pyyntöjä saada oikeus tiettyyn tekijänoikeudellisesti rajattuun aineistoon verkon välityksellä.

”Mahdollisimman pian kaikkien lehtien vuosikerrat I maailmansotaan asti. – sota-aikaan pääseminen tuottaa paljon tutkimusideoita jatkossa, etenkin ruohonjuuritasolla.” (Hölttä, 2016)

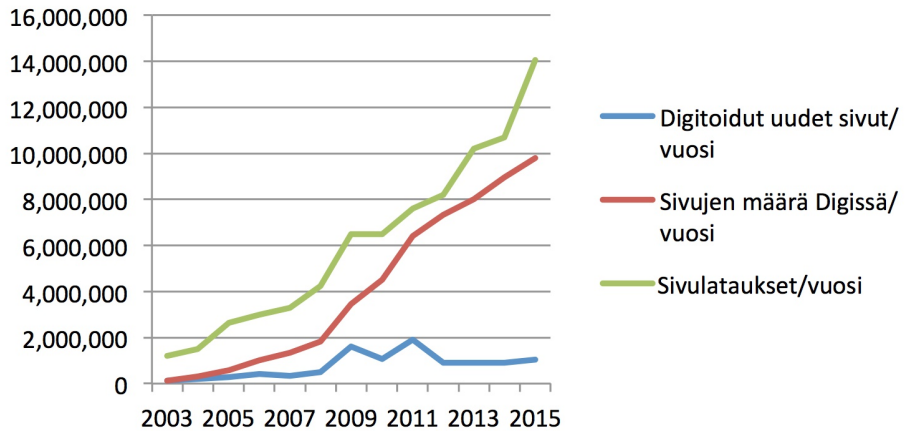
Painettu lehtiaineisto on alkuperäinen aineisto, mutta sen käyttäminen vaatii yleensä matkustamista esimerkiksi vapaakappalekirjastoon tai Kansalliskirjastoon, joten sen saatavuudessa on rajoituksensa. Höltän tutkimuksen vastaajista 84 % ilmoitti mieluummin käyttävänsä digitaalista aineistoa kuin alkuperäistä. Digitaalisen aineistojen käyttämisen perusteluissa mainittiin erityisesti käytön helppous, saavutettavuus ajasta, paikasta ja päätelaitteesta riippumatta, rajoittamaton käyttöaika ja löydettävyys.

Höltän kyselyssä OCR-virheistä mainittiin vain että ne aiheuttavat hämmennystä ja vaikeuttavat hakuja. Digin käyttäjäpalautteiden puolelta on nähtävissä tarkempia ongelmien kuvauksia, kuten seuraavat kaksi esimerkkiä osoittavat.

> ”Ison P-merkin tunnistamisessa fraktuuratekstissä saattaisi olla joku tekninen ongelma, sillä haut eivät palauta tuolla nimellä mitään, mutta nimi löytyy kyllä käyttämällä "Wiipuria”

> ”Hakusanalla ”Lumière” saan kaiken mikä liittyy lumeen”

Näitä ongelmia pyritään ratkaisemaan Kansalliskirjaston palvelussa sillä, että olemme tehneet taustajärjestelmän fraktuurakorjauksiin muutoksia. Yksi vaihtoehto olisi ollut diakriittisten merkkien eli tarkkeiden (é, è jne.) poistaminen, mutta suomen kielessä tämä olisi vienyt pois myös tärkeät skandinaaviset merkit, joten indeksointia on muutettu niin, että se toimii ni-



Kuva 3: Aineistojen digitointi, sivujen määrän kasvu ja sivulataukset digi.fi:ssä

mekkeen pääkielellä. Lisäksi suunnittelemme myös joidenkin yleisimpien OCR-virheiden korjausta taustalla, jotta mahdollisesti puuttuvat hakutulokset saadaan esille ja käytettäväksi.

Digitaalisten aineistojen käyttö näyttää olevan kasvussa, kun sisällytetään verkkokäyttöön eri hakupalvelujen indeksoinnit, henkilö- ja yhteisökäyttäjät ja Kansalliskirjaston oma käyttö eri tarkoituksiin. Verkko-osumien määrään on syytä suhtautua kriittisesti, koska eri käyttäjien aktiivisuudesta riippuen määrät voivat muuttua runsaastikin eri kuukausien välillä. Tutkimus- ja opetuskäyttö luonnollisesti vähenee kesäkuukausina, jolloin taas sukututkimuskäyttö voi lisääntyä. Kuvassa 3 on esitetty digi.kansalliskirjasto.fi:n aineistojen digitoitujen sivujen määrän kasvu, sivujen määrän karttuminen hakupalvelussa sekä sivulatausten vuosittainen kasvu vuosina 2003–2015.

Digin laatuarvio

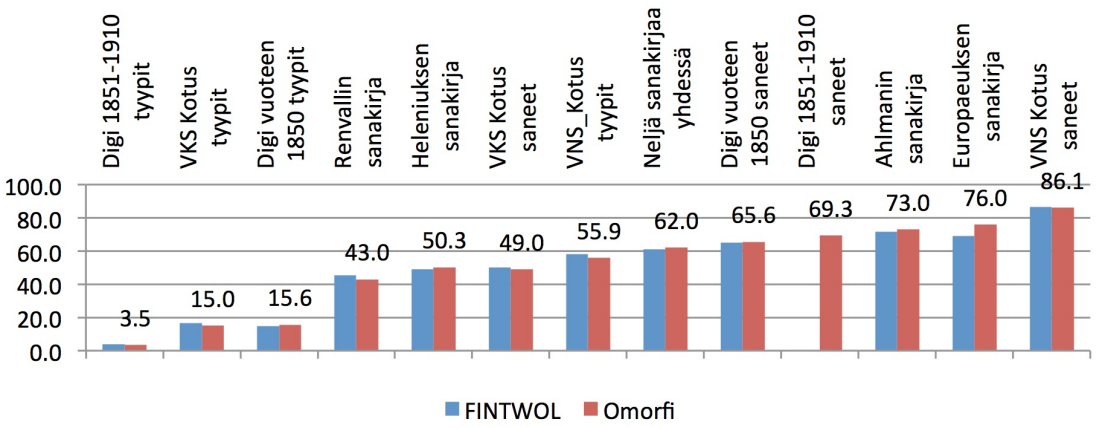
Kansalliskirjaston lehtiaineiston laatua ei ole aiemmin arvioitu systemaattisesti, on vain tiedetty, että laatu ei paikoittain ole kovinkaan hyvä. Vuonna 2014 laadun arviointi aloitettiin otoksilla suomenkielisen aineiston osalta. Saatavilla olevia Kotimaisten kielten keskuksen käsin editoituja 1800-luvun sanomalehtiaineistoja verrokkiaineistoina käyttäen muodostettiin seitsemän pientä rinnakkaiskorpusta, joissa oli yhteensä noin 212 000 sanaa (Kettunen ja kumppanit, 2014; Kettunen, 2016). Parhaimmissa digitoiduissa aineistoissa sanojen oikeellisuus oli noin 90 %, huonoimmassa alle 60 %, ja keskiarvona oli noin 75 %. Luodut rinnakkaiskorpukset olivat kuitenkin äärimmäisen pieni osa lehtiaineistosta ja niiden perusteella tehtävä laatuarvio on välttämättä vajavainen. Laajempien rinnakkaiskorpusten muodostamiseen ei myöskään ole mahdollisuutta: valmista verrokkiaineistoa ei ole saatavilla juuri enempää (Lauerma, 2012) ja resurssit eivät riitä uusien aineistojen käsityönä luomiseen.

Vaihtoehtoja kattavan laatuarvion tekemiseen Digin koko aineistosta ei ole monta, ja suurin osa niistä ei soveltunut käyttöön. Tarkistaminen ihmistyönä ei tule kysymykseen näin laajassa aineistossa, ei edes joukkoistamalla, koska suomen kielen osaajien väestöpohja ei riitä. Riittävän kokoisen luotettavan otokseen perustuvan rinnakkaisaineiston luominen olisi vaatinut myös paljon työtä (vrt. Tanner ja kumppanit, 2009).

Niinpä päädyimme käyttämään apuna kieliteknologiaa. Suomenkielisen lehtikokoelman indeksin sanat analysoitiin alkuun kahdella nykysuomen morfologisella tunnistusohjelmalla (Omorfi ja FINTWOL) ja tunnistettujen ja tunnistamattomien sanojen määrän perusteella arvioitiin indeksin sanaston kokonaislaatu. Tulos on arvio, jossa voi olla usean prosenttiyksikön virhe, mutta se antaa riittävän tarkan kuvan ai-

Taulukko 1: Suomenkielisen sana-aineiston tunnistusprosentit

Kokoelma	**Sanamäärä**	**Omorfi 0.1:n tunnistamia**	**FINTWOLin tunnistamia**
Digin suomenkieliset sanat vuoteen 1850 sananmuodot	22.8 M	65.6 %	65.2 %
Digin suomenkieliset sanat 1851–1910 sananmuodot	2.385 G	69.3 %	---
Digin suomenkieliset sanat vuoteen 1850 sanatyypit	3.24 M	15.6 %	14.9 %
Digin suomenkieliset sanat 1851–1910 sanatyypit	177.3 M	3.8 %	3.5 %



Kuva 4: Digin ja 1800-luvun verrokkiaineistojen tunnistusprosentit, luvut OMorfi 0.1:n tuloksia

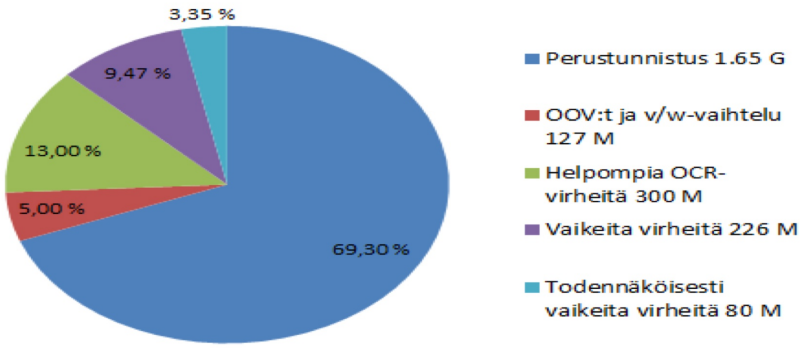
neiston kokonaislaadusta. Sanojen tunnistetuksi tuleminen ei luonnollisesti merkitse sitä, että sana olisi juuri se, mikä se oli alkuperäisessä dokumentissa, mutta todennäköisesti sana on kuitenkin oikein. Käytetystä prosessista syntyi samalla menetelmä, jolla aineiston laatua voidaan arvioida uudestaan, jos aineistoa korjataan esimerkiksi uudella optisella merkintunnistuksella tai ohjelmallisella korjauksella.

Taulukossa 1. on esitetty aineistojen sanojen yleinen tunnistaminen. Aineisto on jaettu kahteen osaan: vuosiin 1771–1850 ja vuosiin 1851–1910.

Saadaksemme käsityksen saman aikakauden sanojen yleisestä tunnistamisesta nykysuomen morfologisilla analyysiohjelmilla tarvitsimme verrokkiaineistoa. Sitä oli saatavilla Kotimaisten kielten keskuksen aineistoista. Kuvassa 4 näytetään lehtiaineiston sanojen tunnistus verrattuna Kotimaisten kielten keskuksen saatavilla oleviin editoituihin 1800-luvun kieliaineistoihin. Luvut viittaavat Omorfin tunnistamien sanojen prosenttiosuuksiin. Kuvassa on kahdenlaisia sanalistauksia: tyyppi- ja sanetason listauksia. Tyyppitasolla kutakin teksteissä esiintyvää sananmuotoa edustaa yksi muoto, sanetasolla ovat mukana kaikki tekstin sanamuodot (saneet).

Kuvasta voi päätellä useita asioita. Tärkein on se, että nykysuomen morfologiset tunnistimet tunnistavat 1800-luvun myöhemmän vaiheen suomea kohtuullisen hyvin, joten niitä voi käyttää laatuarviossa. Varhempaa aineistoa (esimerkiksi VKS, Renvallin ja Heleniuksen sanakirjat) tunnistetaan selkeästi huonommin. Parhaiten on tunnistettu sanat Kotuksen VNS-korpuksesta, jossa on vajaat 5 miljoonaa sanetta. Digin laajemmasta osuudesta, vuosien 1851–1910 sanastosta, tunnistuu noin 69 prosenttia. Sanamäärinä tämä merkitsee sitä, että noin 1.650 mrd sanaa tunnistetaan ja noin 625 miljoonaa jää tunnistamatta.

Tämä on yleiskuva aineistosta, mutta voimme tarkastella aineistoa vielä yksityiskohtaisemmin. Suurten korpusten analyyseista tiedetään, että aineistojen yleisimmät sanat ovat useammin oikein kuin harvinaiset sanat (Ringlstetter ja kumppanit, 2006). Tämä pätee myös Digin aineistossa. Sen miljoona yleisintä sanatyyppiä muodostaa esiintymätasolla 85.6 % koko aineistosta. Tästä aineistosta FINTWOL tunnistaa 79.1 %. Aineiston harvinainen osuus, erityisesti vain kerran aineistossa esiintyvät sanat, tulee hyvin harvoin tunnistetuksi: vain noin 2 % kerran aineistossa esiintyvistä sanoista tunnistetaan. Kymmenen kertaa aineistossa esiintyvistä sanoista vain noin 14 % tunnistetaan. Erityisesti harvaan, 1–10 kertaa aineistossa esiintyvä sanajoukko, noin 225 miljoonaa sanaa, on mitä todennäköisimmin vaikeita optisen merkintunnistuksen aiheuttamia virheitä (Kettunen ja Pääkkönen, 2016).



Kuva 5: Digin 1851–1910 sanaston arvioitu laatu

Muutama kommentti aineiston erityislaadusta on paikallaan. 1800-luvun suomen oikeinkirjoitus alkoi olla jo melko vakiintunutta. Yksi silmiinpistävistä eroista nykysuomeen verrattuna on runsas w:n käyttö v:n sijasta (owat:ovat). Nykykielen morfologiset analyysiohjelmat eivät yleensä tunnista sanoja, joissa esiintyy w, ellei kyseessä ole nimi (Wien, Wagner). Tämän ilmiön vaikutuksen arvioimiseksi muutimme aineistossa w:t v:iksi ja analysoimme aineistot myös näin. Analyysin mukaan 1 miljoonan yleisimmän sanatyypin 2.043 miljardin saneen 427 miljoonasta tunnistamattomasta saneesta tunnistettiin 52 miljoonaa sanaa lisää. V/w-vaihtelulla on siis huomattava vaikutus tunnistamattomiin sanoihin. Kotimaisten kielten keskuksen aineistoissa ero ei näy samalla lailla, koska aineiston toimitusperiaatteiden mukaisesti suuri osa w:istä on muutettu v:iksi.

Toinen tunnistamiseen vaikuttava tekijä on morfologisten ohjelmien sanasto, joka ei voi koskaan olla kattava. Osa tunnistamatta jäävistä sanoista onkin ns. OOV:itä (out-of-vocabulary), sanaston ulkopuolisia sanoja. Näiden määrästä voidaan esittää vain arvioita, mutta se voisi olla 10–20 % (Kettunen ja Pääkkönen, 2016).

Kuvassa 5 on esitetty kokonaisanalyysin lopputulos, jossa on otettu huomioon sekä v/w-vaihtelu että OOV-sanat ja arvioitu virheiden vaikeusasteita.

Arviolta noin 74–75 % Digin vuosien 1851–1910 aineiston sanoista voidaan siis tunnistaa ja noin neljännes sanoista on tunnistamattomia.

Tähän asti esitellyt tulokset ovat analyysimme alkuvaiheesta. Olemme analysoineet aineistoa myös Omorfin versiolla 0.2 (julkaistu 2014) ja versiolla, jota on muokattu tunnistamaan

Taulukko 2: Tunnistustulokset Omorfi 0.2:lla ja HisOmorfilla

Kokoelma	**Sanamäärä**	**Omorfi 0.2**	**HisOmorfi**
Digin suomenkieliset sanat vuoteen 1850 sananmuodot	22.8 M	66.3 %	70.8 %
Digin suomenkieliset sanat 1851–1910 sananmuodot	2.385 G	69.7 %	72.7 %
Digin suomenkieliset sanat vuoteen 1850 sanatyypit	3.24 M	16.0 %	19.4 %
Digin suomenkieliset sanat 1851–1910 sanatyypit	177.3 M	3.9 %	4.9 %

1800-luvun suomen erityispiirteitä. Kutsumme tätä versiota HisOmorfiksi. Taulukossa 2 esitetään näiden kahden ohjelman tunnistustulokset.

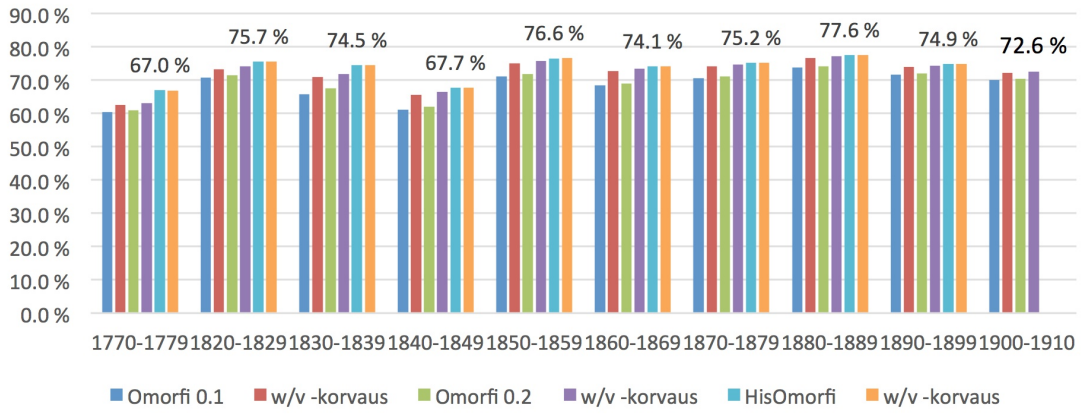
Taulukosta ilmenee, että Omorfin versio 0.2 tunnistaa sanoja vain hiukan paremmin kuin versio 0.1 (vrt. taulukko 1). HisOmorfi tunnistaa sanoja kolme prosenttiyksikköä paremmin kuin Omorfi 0.1 ja FINTWOL. Tämä johtuu pääsääntöisesti siitä, että HisOmorfi käsittelee w:n oikein (vrt. myös kuva 6).

Aineiston vuosikymmenittäinen tunnistus on esitetty kuvassa 6 kolmella eri Omorfin versiolla. Pylväiden yllä oleva luku kertoo HisOmorfin tunnistustuloksen muuten, mutta viimeiseltä ajanjaksolta tulosta ei saatu aineiston merkkiongelmien vuoksi ja tässä tulos on Omorfi 0.2:lla ja w/v-korvauksella.

Kuvasta näkee, että sanojen tunnistus on melko tasaista eri vuosikymmenien aineistoissa. Merkille pantavaa on muutama asia. Ensinnäkin voisi olettaa, että 1800-luvun alkupään aineistoa tunnistettaisiin huonommin kuin vuosisadan loppupään aineistoa (vrt. kuvan 4 tuloksia eri ajanjaksojen aineistojen tunnistamisesta). Erot eivät kuitenkaan ole niin suuria. Tässä voi näkyä lehtien painotavan muutos: 1800-luvun alussa painettiin paljolti yksipalstaisia lehtiä, joissa käytettiin suurta kirjasinlajia. Vuosisadan lopulla käytössä oli usein 5–8 palstaa ja käytetty kirjasin oli pientä. On mahdollista, että painotapa on vaikuttanut sanojen tunnistukseen: vaikka 1800-luvun alun tekstiä tulisi tunnistaa sanastoltaan huonommin kuin myöhempää tekstiä, teksti on ollut helpommin tunnistettavaa optiselle luvulle kuin monipalstainen ja pienikirjasiminen myöhempi aines. Kaksi erilaista vaikuttavaa tekijää tasoittaa mahdollisesti eri vuosikymmenien sanojen tunnistamista. Myös aineistojen sanamäärällä voi olla vaikutusta.

Tässä vaiheessa on syytä käyttää tarkentava metodinen puheenvuoro käytetystä analyysimenetelmästä. Olemme olettaneet, että voimme käyttää laatuanalyysissa nykysuomelle tehtyjä morfologisia analyysiohjelmistoja. Kun aloitimme analyysit, muuta ei ollut saatavillakaan – HisOmorfi tuli saataville vasta myöhemmin. Olemme pyrkineet osoittamaan, että analyysit ovat mielekkäitä rinnastamalla digitoidun aineiston analyysituloksia käsin editoidun vastaavan ajan aineiston analyyseihin. Ne näyttävät melko yhteneviltä. Toinen huomattava asia on luonnollisesti se, että morfologisen ohjelman antama tunnistus sanalle ei merkitse vielä sitä, että sana olisi oikein, eli juuri se sana, joka on ollut lehtitekstissä. Automaattinen morfologinen analyysi voi tehdä monenlaisia vääriä tulkintoja. Listauksessa 1 on vain muutamia analyyseista poimittuja.

Esimerkeissä on useita tyypillisiä tunnistuksen ongelmia: optisessa luvussa on syntynyt paljon katkenneita sanoja, ja niiden osia voidaan



Kuva 6: Omorfin eri versioiden tunnistamien sanojen prosenttiosuus vuosikymmenittäin

Listaus 1: Morfologisen analyysin väärintulkintoja

- mli Num Roman Nom Sg → ilmeinen optisen luvun virhe, tunnistettu silti

- huu huu Part
  tain tai N Gen Sg → sana on jakautunut tavutuksen vuoksi väärin, ja molemmat erilliset osat on tunnistettu (*huutain* olisi käypää 1800-luvun suomea, mutta ei tulisi tunnistetuksi)

- Hei He Pron Nom Pl → tavutus on jakanut sanan, pitäisi olla *heidan*, tunnistamaton
  dan +?

- Samoinkuin +? → kirjoitettu yhteen, jäänyt tunnistamatta
- ylöskannetaan +? → kirjoitettu yhteen, jäänyt tunnistamatta

tunnistaa, vaikka se ei olisikaan mielekästä. 1800-luvulla yhdyssanojen kirjoitustapa oli toisenlaista, mikä ilmenee kahdesta viimeisestä esimerkistä. Omorfin sanasto on myös hyvin laaja, Pirisen (2015b) mukaan sen sanakirjassa on 424 259 lekseemiä. Tämä saa ohjelman tunnistamaan paljon olemattomia yhdyssanoja, jos sanojen osat löytyvät sanakirjasta. Näiden ilmiöiden määrän arvioiminen koko aineistossa on melko mahdotonta, eikä voida sanoa, kuinka paljon ne vaikuttavat tuloksiin.

Toinen huomattava metodinen seikka analyyseissa on, että lehtiaineistojen verrokkeina käyttämämme saman aikakauden editoidut aineistot ovat kovin erikokoisia kuin lehtiaineistomme. Aineistojen kokoerot voivat tuoda useiden prosenttiyksiköiden eroja analyyseihin (Baayen, 2001; Kilgariff, 2001). Tulokset onkin otettava suuntaa antavina, mutta ne antavat mielestämme kuitenkin selvän pohjan, jonka perusteella aineiston laatua ja siinä tapahtuvia muutoksia voidaan analysoida.

Aineiston korjaaminen

Olemme esittäneet artikkelissa tähän saakka arvioita digitoidun historiallisen sanomalehtiaineiston laadusta. Laadun nykytilan arviointi ei kuitenkaan ole prosessin ainoa päämäärä, vaan tarkoituksemme on samalla luoda menetelmä, jolla aineiston laadun parantumista voidaan mitata.

Aineiston laadun parantamiseen on käytännössä kaksi mahdollisuutta: aineistolle suoritettava uusi optinen luku ja ohjelmallinen jälkikorjaus. Parhaat optiset lukuohjelmat ovat usein kaupallisia ohjelmia, ja on osoittautunut, että aineistoa ei voida lukea uudestaan tuoreimmalla versiolla ohjelmasta, jolla alkuperäinen aineisto on luettu (ABBYY FineReader), koska ohjelman lisenssimaksut ovat liian kalliit. Olemmekin tehneet työtä avoimen lähdekoodin optisen lukuohjelman Tesseractin opettamisessa lukemaan fraktuura-fonttia. Toistaiseksi paras Tesseractilla saavutettu tulos testiaineistolla on hiukan parempi kuin vanhan optisen luvun taso. Vanhan aineiston sanavirheluku (WER, word error rate) on 26.10, Tesseractilla saavutettu paras tulos on 25.08. Ero ei ole suuri, ja sen kasvattaminen useaan prosenttiyksikköön vaatii vielä paljon työtä.

Toinen mahdollisuus laadun parantamiseen on ohjelmallinen jälkikorjaus. Kokeilimme jälkikorjausta pienillä aineistoilla ja yksinkertaisella menetelmällä vuonna 2014 (Kettunen, 2016), mutta tulokset eivät olleet riittävän hyviä. Parempia tuloksia olemme saavuttaneet yhteistyössä FIN-CLARIN-konsortion kanssa: heidän jälkikorjausohjelmansa yksi versio on

tuottanut noin yhdeksän prosenttiyksikön parannuksen sanojen tunnistettavuuteen (Kettunen, Pääkkönen, Koistinen, 2016). Kokeilut jatkuvat edelleen, parempi versio jälkikorjausohjelmasta on saatu aikaiseksi FIN-CLARINissa (Silfverberg, Kauppinen, Linden, 2016).

Tässä vaiheessa aineiston laadun parantumisen lopullista tavoitetta on vaikea asettaa. Selvää on, että aineistossa on erittäin paljon vaikeita virheitä, joita ei voida korjata kuin uudelleen skannaamalla. Tähän puolestaan ei ole laajasti mahdollisuutta. Erilaisia jälkikäsittelyitä voidaan kuitenkin yrittää tuotantojärjestelmässä. Toiveissa on, että uudella optisella luvulla ja jälkikorjauksella aineiston sanojen tunnistettavuus nousisi yli 80 prosenttiin. Laatu ei senkään jälkeen olisi kovin hyvä (vrt. Holley, 2009), mutta jo tämän vertainen laadun paraneminen parantaisi aineiston käytettävyyttä.

Kiitokset

Tutkimus- ja kehitystyötä on rahoitettu Euroopan Unionin aluekehitysrahaston Vipuvoimaa EU:lta 2014–2020 –ohjelmasta.

Lähteet

Alex, B., Burns, J. (2014). Estimating and Rating the Quality of Optically Character Recognised Text. In: DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 97–102. http://dl.acm.org/citation.cfm?id=2595214.

Dunning, Alastair (2012). European Newspaper Survey Report. http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf

Hitchcock, T. (2013). Confronting the Digital, Or How Academic History Writing London the Plot. Cultural and Social History, 10:1, 9–23.

Holley, R. (2009). How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine, 15(3/4). http://www.dlib.org/dlib/march09/holley/03holley.html.

Hölttä, T. (2016). Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat Pro gradu, Informaatiotutkimuksen ja interaktiivisen median tutkinto-ohjelma. https://tampub.uta.fi/handle/10024/98714

Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M., Kettunen, K. (2015). Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. Journal of the Association for Information Science and Technology, doi: http://onlinelibrary.wiley.com/doi/10.1002/asi.23379/epdf

Kantor, P. B., Voorhees, E. M. (2000). The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Texts. Information Retrieval,2, 165–176.

Kettunen, K. (2016). Keep, Change or Delete? Setting up a Low Resource OCR Post-correction Framework for a Digitized Old Finnish Newspaper Collection. 11th Italian Research Conference on Digital Libraries - IRCDL 2015. Teoksessa D. Calvanese et al. (Eds.): IRCDL 2015, CCIS 612, 1–9, 2016. DOI: 10.1007/978-3-319-41938-1_11

Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., Kervinen, J. (2014). Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. IFLA World Library and Information Congress, Lyon. http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf

Kettunen, K. Mäkelä, E., Kuokkala, J., Ruokonen, T., Niemi, J. (2016). Modern Tools for Old Content – in Search of Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. LDWA 2016, http://ceur-ws.org/Vol-1670/paper-35.pdf

Kettunen, K., Pääkkönen, T. (2016). Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. LREC 2016 http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf

Kettunen, K., Pääkkönen, T., Koistinen, M. (2016). Between Diachrony and Synchrony: Evaluation of Lexical Quality of a Digitized Historical Finnish Newspaper and Journal Collection with Morphological Analyzers. Baltic HLT, 2016.

Kilgariff, A., (2001). Comparing Corpora. International Journal of Corpus Linguistics 6:1, 97–133.

Klijn, E. (2008). The Current State-of-art in Newspaper Digitization. A Market Perspective. D-Lib Magazin 14(1/2). http://www.dlib.org/dlib/january08/klijn/01klijn.html.

Lauerma, P. (2012). Varhaisnykysuomen korpus ja sen tekstuaalinen edustavuus. Teoksessa Vesa Heikkinen, Eero Voutilainen, Petri Lauerma, Ulla Tiililä & Mikko Lounela (toim.): Genreanalyysi – tekstilajitutkimuksen käytäntöä, s. 308–312. Kotimaisten kielten keskuksen verkkojulkaisuja 29. Helsinki: Kotimaisten kielten keskus. http://scripta.kotus.fi/www/verkkojulkaisut/julk29/

Lopresti, D. (2009). Optical character recognition errors and their effects on natural language processing. International Journal on Document Analysis and Recognition, 12: 141–151.

Mittendorf, E., Schäuble, P. (2000). Information retrieval can cope with many errors. Information Retrieval, 3(3): 189–216.

Nadeau, D., Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. Linguisticae Investigationes 30(1): 3–26.

Niklas, K. (2010). Unsupervised Post-Correction of OCR Errors. Diploma Thesis, Leibniz Universität, Hannover. www.l3s.de/~tahmasebi/Diplomarbeit_-Niklas.pdf

Packer, T., Lutes, J., Stewart, A., Embley, D., Ringger, E., Seppi, K., Jensen, L. S. (2010). Extracting Person Names from Diverse and Noisy OCR Text. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data. Toronto, ON, Canada: ACM.

Piotrowski, Michael. (2012), Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Pirinen, T. (2015a). Omorfi—Free and open source morphological lexical database for Finnish. In Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015 http://www.computing.dcu.ie/~tpirinen/Pirinen-2015-nodalida-omorfi.pdf

Pirinen, T. (2015b). Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. SKY Journal of Linguistics, vol. 28, 381–393. http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Pirinen.pdf

Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., Mäkelä, E. (2016). Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. D-Lib Magazine, July/August.

Ringlstetter, C., Schulz, K., Mihanov, S. (2006). Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. Computational Linguistics 32(3), 295–340.

Savoy, J., Naji, N. (2011). Comparative information retrieval evaluation for scanned documents. Proceedings of the 15th WSEAS international conference on Computers, 527–534.

Silfverberg, M., Kauppinen, P., Linden, K. (2016). Data-Driven Spelling Correction Using Weighted Finite-State Methods. Teoksessa Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata, https://aclweb.org/anthology/W/W16/W16-2406.pdf, 51–59

Strange, C., Wodak, J., Wood, I. (2014). Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers. Digital Humanitites Quarterly, 8. http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html.

Taghva, K., Borsack, J., Condit, A. (1996). Evaluation of Model-Based Retrieval Effectiveness with OCR Text. ACM Transactions on Information Systems, 14(1), 64–93.

Tanner, S., Muñoz, T., Ros, P. H. (2009). Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. D-Lib Magazine, (15/8) http://www.dlib.org/dlib/july09/munoz/07munoz.html.

Traub, M. C., Ossenbruggen, J. van, Hardman, L. (2015). Impact Analysis of OCR Quality on Research Tasks in Digital Archives. Teoksessa Kapidakis, S., Mazurek, C., Werla, M. (toim.), Research and Advanced Technology for Libraries. Lecture Notes in Computer Science, vol. 9316: 252-263.

Aineistolähteet

Ahlmanin sanakirja. Helsinki: Kotimaisten kielten keskus. http://kaino.kotus.fi/korpus/1800/meta/ahlman_sanastot/ahlman_sanastot_coll_rdf.xml.

Europaeuksen sanakirja. Helsinki: Kotimaisten kielten keskus. http://kaino.kotus.fi/korpus/1800/meta/europaeus_sanastot/europaeus_sanastot_coll_rdf.xml.

Heleniuksen sanakirja. Helsinki: Kotimaisten kielten keskus. http://kaino.kotus.fi/korpus/1800/meta/helenius/helenius_coll_rdf.xml.

Renvallin sanakirja. Helsinki: Kotimaisten kielten keskus. http://kaino.kotus.fi/korpus/1800/meta/renvall/renvall_coll_rdf.xml.

VKS, vanhan kirjasuomen taajuuslista. Helsinki: Kotimaisten kielten keskus. http://kaino.kotus.fi/sanat/taajuuslista/vks.php.

VNS, varhaisnykysuomen taajuuslista. Helsinki: Kotimaisten kielten keskus. http://kaino.kotus.fi/sanat/taajuuslista/vns.php.