VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Heli Rantala, Hannu Salmi, Asko Nivala, Petri Paju,
Reetta Sippola, Aleksi Vesanto ja Filip Ginter

Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920

Digitaalisten ihmistieteiden näkökulma

Sanomalehdet ovat historiantutkijoiden keskeistä lähdeaineistoa, ja Suomessa Kansalliskirjaston digitoimat sanoma- ja aikakauslehdet tuovat tämän kattavan materiaalin helposti niin tutkijoiden kuin muiden kiinnostuneiden ulottuville. Suomalaiset laajat lehdistöhistoriat – esimerkiksi Clas Zilliacuksen ja Henrik Knifin *Opinionens tryck* (1985), Päiviö Tommilan toimittama *Suomen lehdistön historia* -sarja (1985–1992) sekä myöhempi Tommilan ja Raimo Salokankaan *Sanomia kaikille* -teos (1998)[1]– on kuitenkin kirjoitettu aikana, jolloin digitoitu lehtimateriaali ei vielä ollut tutkijoiden saatavilla. Tässä artikkelissa tuomme suomalaisen lehdistön tutkimiseen näkökulman, joka ei ole ollut mahdollinen ennen Kansalliskirjaston digitaalista korpusta. Tutkimuskohteenamme on koko lehdistöaineisto, ja sitä lähestytään datatieteissä kehitetyn laskennallisen menetelmän avulla.

Määrällisellä, laskentaan perustuvalla lähestymistavalla on sinänsä pitkät perinteet sanomalehtien tarkastelussa. Vuonna 1973, mittavan *Suomen sanomalehdistön historia* -projektin esivaiheessa, historiantutkija Viljo Rasila esitteli eri mahdollisuuksia käyttää matemaattisia analyysimalleja lehdistötutkimuksessa, ja tutkijat keskustelivat tehtävään parhaiten soveltuvista kvantitatiivisista menetelmistä kuten palstamillimetrien mittaamisesta. Lehtiaineistoa oli kuitenkin vielä pitkään käsiteltävä paperisena ja mikrofilmeiltä lukemalla.[2] Kansalliskirjasto avasi digitoitujen lehtien kokoelmansa vuonna 2001, jolloin tarjolla oli 36 000 sivua.[3] Tänä vuonna Kansalliskirjaston digitaalinen korpus kattaa jo yli 13 miljoonaa avoimessa verkkokäytössä olevaa sivua.

Sittemmin eri vauhdilla edenneitä kansallisia sanomalehtien digitointiprojekteja on ollut käynnissä runsaasti. Vanhojen tekstiaineistojen digitoiminen onkin ollut pitkä prosessi, joka on sisältänyt paljon myös arviointi- ja kehitystyötä.[4] Samalla on pohdittu sitä, miten sanoma- ja aikakauslehtiaineiston digitoiminen muuttaa käsityksiämme lehdistön ja julkisuuden historiasta. Aineiston käytön merkittävän helpottumisen lisäksi erilaisten laskennallisten menetelmien hyödyntäminen on tullut mahdolliseksi digitoinnin myötä. Toisaalta aineiston digitoiminen on nostanut esille myös uudenlaisia kysymyksiä, jotka liittyvät muun muassa digitoinnin laatuun ja historiallisen aineiston säilyttämiseen.[5] Usein

1. Clas Zilliacus & Henrik Knif, *Opinionens tryck. En studie över pressens bildningsskede i Finland.* Svenska litteratursällskapet i Finland 1985; Päiviö Tommila (toim.) *Suomen lehdistön historia* 1–10. Kustannuskiila 1985–1992; Päiviö Tommila & Raimo Salokangas, *Sanomia kaikille. Suomen lehdistön historia.* Edita 1998.
2. Ks. esim. Viljo Rasila, Tilastomatemaattisten analyysien käyttömahdollisuudet lehdistötutkimuksessa. Teoksessa Kristiina Ritari (toim.) *Lehdistöntutkijain seminaari 1973: Alustukset ja keskustelut.* Helsingin yliopiston historian laitoksen julkaisuja N:o 1. Helsingin yliopisto 1973, 112–128. Suomalaisen lehdistö- ja mediahistorian tutkimushistoriasta ks. Kaija Vuorio, *Sanoma, lähettäjä, kulttuuri. Lehdistöhistorian tutkimustraditiot Suomessa ja median rakennemuutos.* Jyväskylän yliopisto 2009, 43–44.
3. Suomalaisen kokoelman syntyhistoriasta, ks. lähemmin Mila Oiva, Hannu Salmi & Asko Nivala, *Digitized Newspapers at the National Library of Finland*, http://oceanicexchanges.org/2018-02-20-data-reports-finland/ (27.6.2018).
4. Kansalliskirjaston osalta ks. esim. Kimmo Kettunen, Tuula Pääkkönen & Mika Koistinen, Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910. Sanatason laatu, kokoelmien käyttö ja laadun parantaminen. *Informaatiotutkimus* 35:3 (2016), 3–14.
5. Mediahistorioitsijat Johan Jarlbrink ja Pelle Snickars ovat jopa esittäneet, että digitoinnissa varsin yleiset koneluennan

laajojen digitaalisten arkistojen materiaalia lähestytään tietyn ennalta määritellyn aihekokonaisuuden tai teeman kautta – esimerkiksi erilaiset sanahaut ovat suosittu tapa selata laajaa aineistoa sopivia osumia etsien. Toisaalta sanahaut eivät periaatteessa poikkea perinteisen kirjan selaamisesta hakemistoa käyttäen: tiedonhaun volyymi vain on huomattavasti suurempi, mutta samalla optiseen tekstin tunnistukseen (OCR, optical character recognition) liittyvät virheet peittävät osan osumista. Entäpä jos tutkimuksessa ei etukäteen lainkaan määriteltäisi, millaisia teemoja aineistosta etsitään, vaan teemat hahmottuisivat ainoastaan menetelmän kautta? Omassa tutkimuksessamme olemme soveltaneet tällaista menetelmälähtöistä lähestymistapaa suomalaiseen lehtiaineistoon.

Tämän artikkelin tavoitteena on tutkia suomalaista sanoma- ja aikakauslehdistöä tekstien uudelleenkäytön näkökulmasta. Uudelleenkäyttö avaa lehdistöön useita mahdollisia analyysin tasoja. Yhtäältä se voi viitata ns. saksijournalismiin, toimintaan, jossa sisältöjä konkreettisesti leikattiin aiemmin ilmestyneistä lehdistä. Toisaalta se voi viitata tekijänoikeudellisiin lähtökohtiin, joissa teksteillä ei ollut oikeudellista suojaa. Lisäksi se voi viitata lehdistön intermediaalisiin suhteisiin: aineistoa siirtyi lehdistöön myös muista lähteistä, lennätinuutisista kiertokirjeisiin. Samalla uudelleenkäyttö antaa mahdollisuuden katsoa tekstien liikkeitä yleisemmällä tasolla, merkkien kiertona lehdistön luomassa verkostossa. Tässä artikkelissa lähtökohtana on, että jos pystymme algoritmisesti tunnistamaan korpuksesta samuuksia, eli toistettuja tekstejä tai tekstifragmentteja, lehdistön informaatiovirtoja on mahdollista analysoida kokonaisuutena. Samalla menetelmä antaa vihjeitä aineiston välisistä sisäisistä suhteista tavoilla, jotka eivät muuten ilmenisi tai joita ei esimerkiksi avainsanahaulla voisi tavoittaa. Tätä kehitystyötä olemme tehneet monitieteisessä COMHIS-konsortiossa.[6] Artikkeli rakentaa kokonaiskuvaa siitä, millaista tekstien kopiointi lehdistössä oli vuosina 1771–1920, miten se muuttui ja millaisia vaihteluita toiston rytmeissä tapahtui. Aikarajauksemme ulottuu ensimmäisestä julkaistusta sanomalehdestä vuodesta 1771 vuoteen 1920, johon asti louhintaan tarkoitetut Kansalliskirjaston avaamat datapaketit ulottuvat.[7]

Kierrätys näkökulmana

Tekstin uudelleenkäytön tutkimuksessa sanomalehdistöä lähestytään ei vain tiedon tuottajana ja välittäjänä vaan informaation kierrättäjänä. Lähestymme tässä artikkelissa kierrätystä ennen kaikkea kopioinnin kautta: miten tekstejä tai tekstifragmentteja siirrettiin suoraan uusiin tekstuaalisiin ympäristöihin. Informaatio viittaa siirrettäviin ja kierrätettäviin merkkeihin, tekstien materiaalisiin virtoihin, kun taas käsitettä 'tieto' käytämme informaation järjestymisen ja tulkitsemisen merkityksessä.

Saman tekstin julkaiseminen uudelleen useissa eri yhteyksissä on vanha ilmiö, joka on tunnettu kautta aikojen, ja se on ollut myös lehdistötutkijoiden havaitsema. Jouko Teperi tutki 1970-luvulla sanomalehdistöstä esimerkiksi susien hyökkäyksiä 1800-luvulla ja tunnisti kymmenien vuosien kuluttua uudelleen julkaistuja tekstejä.[8] *Suomen sanomalehdistön historia* -projektin yhteydessä toteutettiin puolestaan pistetutkimus uutisten siirtymisestä lehdestä toiseen. Tarkastelujaksona olivat tammi- ja helmikuut 1848, mikä antoi kuvaa siitä, millaista uutisten kierrätys oli.[9] Huolellisessa lähiluvussa voitiin tunnistaa myös kielirajan yli tapahtunutta kopiointia ja lainaamista, mutta toisaalta kahden kuukauden aikaväli oli rajalli-

virheet johtavat aineiston säilyttämisen sijaan myös sen tuhoamiseen. He viittaavat tällä siihen, että pahimmillaan koneluennan virheet tekevät tekstistä täysin lukukelvottoman. Johan Jarlbrink & Pelle Snickars, Cultural Heritage as Digital Noise. Nineteenth Century Newspapers in the Digital Archive. *Journal of Documentation* 73:6 (2017), 1228–1243.

6. *Computational History and the Transformation of Public Discourse in Finland, 1640–1910* (COMHIS) on Suomen Akatemian rahoittama konsortio ja osa digitaalisten ihmistieteiden tutkimusohjelmaa. Konsortiossa ovat mukana Suomen Kansalliskirjasto, Helsingin yliopisto ja Turun yliopisto.
7. Sanomalehdet on julkaistu datapaketteina Kielipankin kautta. Käytössämme ovat olleet seuraavat paketit: *The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771–1874)*, julkaistu 2011, http://urn.fi/urn:nbn:fi:lb-201505112; *Newspaper and Periodical OCR Corpus of the National Library of Finland (1875–1920)*, julkaistu 2017, http://urn.fi/urn:nbn:fi:lb-201405275.
8. Jouko Teperi, *Sudet Suomen rintamaiden ihmisten uhkana 1800-luvulla.* SHS 1977, esim. 83–84, 154.
9. Päiviö Tommila, Yhdestä lehdestä sanomalehdistöksi 1809–1859. Teoksessa Päiviö Tommila (toim.) *Suomen lehdistön historia* 1. *Sanomalehdistön vaiheet vuoteen 1905.* Kustannuskiila 1988, 205.

nen. Asetelma olisi haasteellisempi esimerkiksi vuonna 1917, jolloin julkaistiin 130 eri sanomalehteä. Vuonna 1848 lehtiä oli vasta kaksitoista. Ennen digitaalista aikakautta tekstien uudelleenkäyttöä ei olekaan voitu tutkia systemaattisesti, sillä kierrätetyn tekstimassan laajuudesta ei ole käytännössä mahdollista saada kokonaiskuvaa tekstejä lukemalla tai erilaisilla otantamateriaaleilla. Lehdistön, ja laajemmin julkaisukulttuurin, tutkimisessa on lisäksi toisinaan korostunut varsin yksilökeskeinen näkökulma aineistoon. Tällöin huomio on kiinnittynyt toimituksellisesti merkittävinä pidettyihin teksteihin ja niiden laatijoihin.[10] Suomessa tämä näkökulma on korostunut etenkin Zilliacuksen ja Knifin teoksessa, jossa lehdistön kehitystä tarkastellaan voimakkaasti yksittäisten toimittajien toiminnan kautta.

Tuoreessa kansainvälisessä tutkimuksessa sanomalehtien harjoittamaa tekstien kierrättämistä on tutkittu digitaalisin menetelmin etenkin yhdysvaltalaisella aineistolla Ryan Cordellin ja David A. Smithin *Viral Texts* -hankkeessa[11] sekä brittiläisissä lehdissä Melodee Bealsin toimesta[12]. Yhdysvaltoihin tai brittiläiseen materiaaliin verrattuna suomalaisessa aineistossa on omat erityisyytensä. Yksi näistä on sanomalehtien yleisesti käyttämä fraktuurakirjasin, joka vaikeuttaa tekstien koneluentaa. Toinen erityisyys on suomalaisen lehdistön monikielisyys; tarkastelujakson materiaalista 62 prosenttia on suomenkielistä ja 37 prosenttia ruotsinkielistä.[13] Pienemmässä mittakaavassa lehtiä on ilmestynyt myös muilla, erityisesti saksan ja venäjän, kielillä. Tämän lisäksi moneen muuhun maahan, Yhdysvallat mukaan lukien, verrattuna lehdistön kasvu tapahtui Suomessa hitaasti. Vielä 1870-luvun alussa Suomessa ilmestyi vain parisenkymmentä sanomalehteä, kun vuonna 1920 niitä oli jo 136. Se, että Suomessa oli pitkään melko vähän lehtiä, on luonnollisesti rajoittanut mahdollisuuksia lehtien välisiin nopeisiin lainauksiin. Toisaalta olemme havainneet, että varhaisimpien sanomalehtien materiaalia hyödynnettiin pitkään ja niistä saatettiin ottaa tekstilainauksia vuosikymmenten tai jopa yli sadan vuoden kuluttua.

Tarkastelujakson aikana suomalaisessa lehdistössä ja yhteiskunnassa tapahtui suuria muutoksia. Ensimmäinen ”suomalainen” lehti, vuonna 1771 käynnistetty *Tidningar Utgifne af ett Sällskap i Åbo*, tulee nähdä myös osana lehdistön nousua Ruotsin valtakunnassa.[14] Vuodesta 1810 lähtien *Åbo Allmänna Tidning* oli Suomen suuriruhtinaskunnan ainoa sanomalehti, jonka tehtävänä oli jakaa myös viranomaismateriaalia. Lehtien julkaiseminen laajeni hiljalleen Turusta Helsinkiin, Viipuriin ja Ouluun 1820-luvulla ja vähitellen 1800-luvun jälkipuoliskolla[15] myös sisämaan kaupunkeihin. 1900-luvun alkuun mennessä lehtiä ilmestyi lähes 40 paikkakunnalla.

Huomion kiinnittäminen tekstien kopioitumiseen ja lainaamiseen avaa suomalaisen julkaisukulttuurin keskeiseen muotoon, sanoma- ja aikakauslehdistöön, uuden näkökulman. Olemme havainneet uudelleenkäytön olleen hyvin moninaista. Se koostuu sekä erilaisen pysyväisluonteisen materiaalin kuten ilmoitusten ja mainosten, tuomasta toistosta että uutisten, tarinoiden, kertomusten tai runonpätkien lainaamisesta lehdestä toiseen joko nopealla tai hyvinkin pitkällä aikavälillä. Uudelleenkäyttö tarkoittaa tässä tutkimuksessa ennen kaikkea korpuksen sisällä ilmenevää toisteisuutta: mitä tekstejä tai tekstien osia on julkaistu myöhemmin? Ilmeistä on, että julkaisemisen motiivit vaihtelivat. Joukossa on 1) tärkeiksi koettuja uutisia, jotka on haluttu julkaista mahdollisimman nopeasti, 2) mennei-

10. Ryan Cordell, Reprinting, Circulation, and the Network Author in Antebellum Newspapers. *American Literary History* 27:3 (2015), 420; Johan Jarlbrink, Pelle Snickars & Christian Colliander, Maskinläsning. Om massdigitalisering, digitala metoder och svenska dagspress. *Nordicom Information*, 38:3 (2016), 32.
11. Ks. esim. David A. Smith, Ryan Cordell & Elisabeth Maddock Dillon, Infectious Texts. Modelling Text Reuse in Nineteenth-Century Newspapers. *Proceedings of the Workshop on Big Humanities*. IEEE Computer Society Press 2013, 86–94; Cordell 2015, 417–445.
12. Melodee H. Beals, Scissors and Paste. The Georgian Reprints, 1800–1837. *Journal of Open Humanities Data* 3 (2017), https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.8/ (8.6.2018).
13. Vuoteen 1890 saakka ruotsinkielistä materiaalia julkaistiin enemmän kuin suomenkielistä. Kettunen, Pääkkönen & Koistinen 2016, 4.
14. Juuri 1760- ja 1770-luvut olivat Ruotsin valtakunnassa voimakasta lehdistön nousun aikaa. Ks. Claes-Göran Holmberg, Ingemar Oscarsson & Jarl Torbacke, *Den svenska pressens historia* I. *I begynnelsen (tiden före 1830)*. Ekerlids Förlag 2000, 206–207.
15. Kuopio on ainoa sisämaan kaupunki, jossa julkaistiin lehtiä jo ennen 1800-luvun puoliväliä.

syydestä löydettyjä tekstejä, joita on pidetty tärkeänä myös nykypäivälle, 3) kertomuksia, jotka ovat herättäneet innostusta tai suorastaan tartuttaneet lukijansa, mutta myös 4) maksettuja ilmoituksia ja mainoksia, aikatauluja ja hinnastoja sekä viranomaistiedotteita. Aineiston heterogeenisyys on sen etu ja rikkaus: nämä kaikki tekstilajit muodostivat informaatiovirtoja, joiden kautta lehdistön luonnetta verkostona ja järjestelmänä voi tutkia. Toki uudelleenkäytön kohteena olivat myös tekstit, jotka tulivat suomalaisen lehdistön ulkopuolelta: ulkomaisista tiedotuskanavista, kirjeistä, lehdistä ja sähkeistä, tai vaikkapa kirjallisuudesta. Tämän artikkelin analyysi perustuu kuitenkin sanoma- ja aikakauslehtien korpukseen. Materiaali on niin laaja, ettei kopioitujen tekstien alkuperää voi systemaattisesti selvittää. Kerromme seuraavassa tarkemmin itse menetelmästä, jolla käytettävissä olevat viisi miljoonaa sivua on analysoitu. Sen jälkeen tarkastelemme lähemmin uudelleenkäyttöä temaattisesti, aluksi tiedonliikkeitä kokonaisuutena ja sen jälkeen tekstien viraalisuutta ja pitkän aikavälin toistoa. Olemme täydentäneet analyysia tulkitsemalla laadullisesti esimerkkitapauksia alkuperäisten sanomalehtien pohjalta.

Tekstin uudelleenkäytön tunnistus BLAST:lla

Hankkeen lähdeaineistona on Kansalliskirjaston julkaisema sanoma- ja aikakauslehtien digitoitu OCR-korpus, joka on luotu alun perin mahdollistamaan kokotekstihaut digitaalisesta sanomalehtiarkistosta.[16] Kuviksi skannattujen sanoma- ja aikakauslehtien sivut on muutettu tekstiksi OCR:n avulla. OCR-teknologiassa on edelleen puutteita, ja tästä syystä osa merkeistä on tunnistettu väärin, mikä aiheuttaa aineistoon vaihtelevan tasoista kohinaa (*noise*).[17] *Viral Texts* -hankkeessa työskentelevä David A. Smith on kehittänyt tekstien uudelleen käytön tunnistukseen Passim-ohjelman, jota on käytetty menestyksekkäästi yhdysvaltalaisten lehtien tutkimuksessa.[18] Koska Passim perustuu n-grammien käyttöön, se pystyy tietyn kynnyksen puitteissa tunnistamaan tekstien toisinnot samaksi tekstiksi, vaikka tekstit eroaisivat toisistaan joko toimittajan editointityön tai OCR-ongelmien vuoksi.[19] Kokeilimme Passimia myös suomalaisen aineiston analyysiin, mutta suomalaisen korpuksen fraktuurakirjasin yhdessä skannauksen mahdollisen heikkouden ja suomen kielen erityispiirteiden kanssa tekee OCR:stä heikkolaatuisempaa. Passim pohjautuu niin sanottujen perustavien yhtäläisyyksien (*seed overlap*) tunnistamiseen, joiden tulee olla useampien kokonaisten sanojen mittaisia. Tällaiset jaksot ovat valitettavasti varsin harvinaisia meidän aineistossamme, jossa painoajankohdasta riippuen keskimäärin 25–30 % sanoista voi olla väärin tunnistettuja.[20]

Näiden ongelmien vuoksi päädyimme COMHIS-hankkeessa kehittämään oman NCBI BLAST:iin perustuvan ratkaisun, jonka toteuttivat Filip Ginter ja Aleksi Vesanto. BLAST (*Basic Local Alignment Search Tool*) on tarkoitettu biologisten sekvenssien kuten aminohappojen tunnistamiseen ja vertailuun, joten se soveltuu sinänsä erinomaisesti runsaasti kohinaa sisältävien suurten merkkiaineistojen tutkimiseen. BLAST on kuitenkin räätälöity aminohappojen tutkimukseen, joten sen soveltaminen edellytti

16. Lähemmin korpuksesta, ks. Tuula Pääkkönen, Jukka Kervinen, Asko Nivala, Kimmo Kettunen & Eetu Mäkelä, Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. *D-Lib Magazine* 22 (2016), http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html (5.6.2018). Ks. myös viite 7.
17. OCR:n tason parantaminen on myös eräs COMHIS-konsortion tavoite. Ks. Mika Koistinen, Kimmo Kettunen & Tuula Pääkkönen, Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. *Nordic Conference on Computational Linguistics, NoDaLiDa 2017 Gothenburg, Sweden*. Linköping Electronic Conference Proceedings 2017, 277–283.
18. Viral Texts -hankkeen tuloksista, ks. Smith, Cordell & Maddock Dillon 2013; Cordell 2015. Passim on vapaan lähdekoodin ohjelmisto, joka on julkaistu GitHubissa, https://github.com/dasmiq/passim.
19. N-grammilla tarkoitetaan n merkin tai sanan mittaista jaksoa eli esimerkiksi 5-grammi tarkoittaa viiden merkin tai sanan jaksoa. N-grammimalleja käytetään kielitieteessä ja kieliteknologiassa tilastollisten kielimallien rakentamiseen. Passim perustuu sanojen n-grammien vastaavuuksien vertailuun, mutta laskentatehon säästämiseksi se ei vertaa kaikkia hakupareja keskenään. David A. Smith, Ryan Cordell & Abby Mullen, Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History* 27:3 (2015), E1–E15, https://doi.org/10.1093/alh/ajv029 (5.6.2018).
20. Ks. Kimmo Kettunen, Tuula Pääkkönen & Mika Koistinen, Between diachrony and synchrony. Evaluation of lexical quality of a digitized historical finnish newspaper and journal collection with morphological analyzers. *Proceedings of the 7th International Conference: Human Language Technologies – The Baltic Perspective.* Riga, Latvia, October 2016, http://ebooks.iospress.nl/ISBN/978-1-61499-701-6 (26.6.2018).

aineiston esiprosessointia niin että kirjaimet koodattiin 23 aminohapon mukaan. Nämä eivät riitä koko aakkoston kääntämiselle suomen kielessä, jonka takia vain 23 käytetyintä kirjainta käännetään. Muutamia pois jääviä kirjaimia käytetään niin harvoin muihin verrattuna, että ne eivät vaikuta BLAST:in tuloksiin merkittävästi. Tämä poisti samalla aineistosta analyysin ajaksi myös numerot ja erikoismerkit.[21] BLAST mahdollistaa kaikkien sanomalehtisivujen vertailun keskenään niin, että tietyn samankaltaisuuden kynnyksen ylittävät tekstit ryhmitellään samaan klusteriin eli ryppääseen. Asetimme osumien kynnyksen siten, että tunnistettavan tekstijakson on oltava vähintään 300 merkkiä pitkä, jotta välttäisimme sanomalehdille tyypillisen niin sanotun *boiler plate* -tekstin (eli pohjatekstin) tunnistamisen, joka on lehtiin painettua toisteista ja kaavamaista informaatiota ilman varsinaista sisältöä. Tekemämme vertailun perusteella BLAST antaa huomattavasti tarkempia tuloksia kuin *Viral Texts* -hankkeen kehittämä Passim, mutta se kuluttaa samalla myös enemmän prosessoritunteja. Esimerkiksi kokeiluvaiheessa tehty vuosien 1771–1910 aineiston tunnistus kulutti n. 150 000 CPU-tuntia, mikä vastaa arviolta 4 vuoden laskentatyötä kotitietokoneella. BLAST vaatii siis käytännössä tieteelliseen laskentaan erikoistuneen laitteiston käyttämistä.[22]

BLAST-ajon tuloksena saimme tietokannan, jossa osumat (*hit*) eli toisiaan riittävästi muistuttavat tekstijonot on ryhmitelty samaan ryppääseen eli klusteriin (*cluster*). Vuosien 1771–1920 aineistosta saimme lähes 61 miljoonaa osumaa, jotka jakaantuvat 13,8 miljoonaan klusteriin. Olemme verranneet menetelmäämme samaan tarkoitukseen kehitettyyn Passim-ohjelmaan: BLAST löytää selkeästi enemmän toistettuja tekstejä.[23] Tuloksistamme ei kuitenkaan voida suoraan päätellä jaettujen tekstien määrää. Emme ole ensinnäkään tutkineet BLAST:n saantia (*recall*) eli sitä kuinka monta uudelleenkäytettyä tekstiä BLAST kykenee löytämään. Tämä edellyttäisi vertailukorpuksen manuaalista kokoamista, joka on hyvin työlästä. Eräs jatkotutkimuksen aihe olisikin verrata BLAST:n tuloksia aiempaan pistetutkimukseen uutisten kulusta lehdestä toiseen tammi-helmikuussa 1848. Toisaalta OCR:n taso vaihtelee joka tapauksessa paljon eri vuosikymmenillä, joten yksi pistetutkimus voi antaa vain suuntaa-antavaa tietoa.[24] Toiseksi Kansalliskirjaston sanoma- ja aikakauslehtien OCR-korpus on myös jaettu lehtisivuittain eikä artikkeleittain. Tämän vuoksi on mahdollista, että yksi artikkeli on voinut pilkkoutua useampaan klusteriin esimerkiksi sivu- tai palstarajojen takia. BLAST on voinut myös niputtaa esimerkiksi juna-aikatauluja, torihinnastoja ja muita mainoksia sekä ilmoituksia yhteen, jos niissä toistuvien samankaltaisten merkkien määrä on ylittänyt 300 merkkiä. Näitä ongelmia ei voida kuitenkaan välttää, mikäli BLAST:n halutaan tunnistavan myös mahdollisimman tarkasti lehtien toisiltaan kopioimat uutiset, tarinat ja muut sitaatit.

On selvää, että tällaisen tietokannan selaaminen edellyttää varsin edistyneitä haku- ja suodatustoimintoja. Kaikille käyttäjille avoin hakuportaalimme osoitteessa comhis.fi on toteutettu Apache Solr -indeksoinnilla.[25] Hakutoiminnot mahdollistavat haun sekä yksittäisistä osumista että useampien tekstien klustereista. Jokainen klusteri on numeroitu yksilöllisellä tunnisteella. Solr:n filtterien avulla hakuja on mahdollista kohdistaa samanaikaisesti esimerkiksi tiettyyn

21. Esimerkiksi merkkijono "Tämä on esimerkkilause." olisi aminohapoiksi koodattuna "DWNWIEGHCNGPMMCFBKHG". BLAST:n toiminnasta, ks. Aleksi Vesanto, Asko Nivala, Tapio Salakoski, Hannu Salmi & Filip Ginter, A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora. *Proceedings of the 21st Nordic Conference of Computational Linguistics.* Gothenburg, Sweden, 23–24 May 2017 (Linköping 2017), 330–333, http://www.ep.liu.se/ecp/131/049/ecp17131049.pdf (27.6.2018); Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers & David J. Lipman, Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215 (1990), 403–410.

22. Aleksi Vesanto, Asko Nivala, Heli Rantala, Tapio Salakoski, Hannu Salmi & Filip Ginter, Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910. *Proceedings of the 21st Nordic Conference of Computational Linguistics. Gothenburg,* Sweden, 23–24 May 2017. Linköping Electronic Conference Proceedings 2017, 54–58, http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf (5.6.2018). Hankkeen laskentaresurssit tarjosi Espoossa toimiva CSC – Tieteen tietotekniikan keskus. Ks. myös Hannu Salmi, Viraalisuus. Kulttuurihistoriallinen näkökulma. *niin & näin* 1 (2018), 74–75.

23. Ibid.

24. Ks. Tommila 1988, 205.

25. Aleksi Vesanto, Filip Ginter, Hannu Salmi, Asko Nivala, Reetta Sippola, Heli Rantala & Petri Paju, *Text Reuse in Finnish Newspapers and Journals*, 1771–1920, http://comhis.fi/clusters (6.3.2019).

yksittäiseen sanomalehteen, tietyn kielisiin lehtiin, pelkästään sanomalehtiin tai aikakauslehtiin, tiettyyn aikahaarukkaan tai maantieteelliseen alueeseen. Jokainen osuma sisältää myös linkin kyseisen lehden sivuun Kansalliskirjaston digitaalisessa sanoma- ja aikakauslehtien portaalissa, jolloin käyttäjä voi halutessaan lukea samaa tekstiä alkuperäisenä skannauksena. Solr pystyy myös sumeaan hakuun (*fuzzy search*), joka etsii hakutermin lisäksi muita samankaltaisia termejä. Tämä on hyödyllinen ominaisuus OCR-ongelmien takia, sillä sanoissa on usein virheitä. Hakuportaalimme sisältää myös analyysityökaluja, joilla voi tarkastella Solriin tehdyn haun perusteella saatujen tulosten ajallista sekä maantieteellistä kattavuutta.[26]

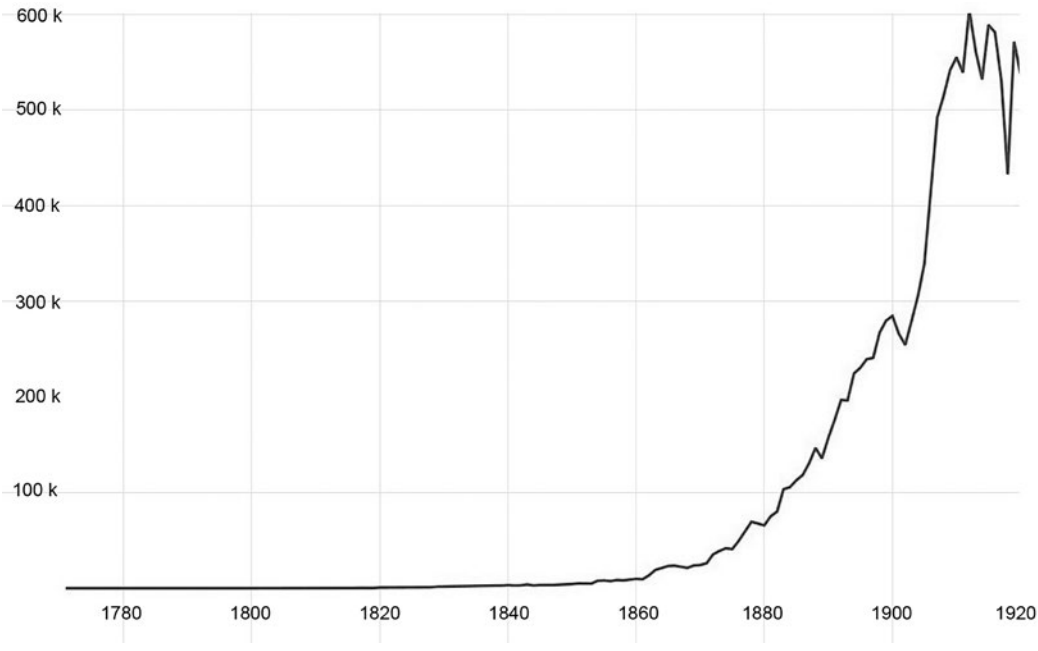
Jokaiselle klusterille on myös laskettu arvo, joka pyrkii arvioimaan tämän ”viraalisuutta” (*virality score*). Hyödynnämme tätä analyysityökalua myöhemmin artikkelin alaluvussa ”Viraalisuuden jäljillä”. Arvoa varten lasketaan, kuinka moneen uniikkiin lehteen ja paikkakuntaan uutinen on levinnyt sekä kuinka monta päivää tämä kesti. Arvo saadaan kertomalla lehtien ja paikkakuntien määrä kuluneen ajan käänteisluvulla. Tämä siis rankaisee arvoa, jos uutinen ei ole levinnyt tarpeeksi tai siinä on kestänyt liian kauan. Laskuista jätetään myös pois klustereiden osumat, jotka selvästi poikkeavat muiden osumien päivämääristä, jotta laskettu arvo ei vääristyisi. Tämä voi tapahtua esimerkiksi tilanteessa, jossa uutinen on levinnyt nopeasti lyhyellä aikavälillä, mutta siitä on julkaistu yksittäinen teksti paljon myöhemmin. Tällöin klusterin aikajänne (*span*) on pitkä, mutta tosiasiassa leviäminen on tapahtunut nopeasti. Lopuksi kaikkien klustereiden arvot normalisoidaan nollan ja sadan välille selkeyden vuoksi. Viraalisuusarvojen tulkinnassa sanoma- ja aikakauslehtien erot on syytä ottaa huomioon. Käytetyt metatiedot ovat peräisin Kansalliskirjastolta ja ne ovat pääosin hyvin luotettavia, mutta metatiedoissa voi esiintyä yksittäisiä virheitä. Aikakauslehtien metatiedoissa julkaisupäivää ei ole voitu merkitä samalla tarkkuudella kuin sanomalehtien kohdalla, mikä vaikuttaa tuloksiin. Kokonaiskuvaa tämä ei kuitenkaan muuta, sillä levinneimmät tekstit löytyvät sanomalehdistöstä. Yksittäisen klusterin viraalisuusarvon tulkinnassa tämä täytyy kuitenkin ottaa huomioon.

Tiedon liike uudelleenkäytön näkökulmasta

Hankkeessa sanoma- ja aikakauslehtikorpuksesta on löytynyt BLAST-työkalun avulla toistoa sisältäviä tekstejä tai tekstikatkelmia noin 13,8 miljoonan klusterin verran. Kuvassa 1 on esitetty klustereiden määrän kehitys aikavälillä 1771–1920. Kuvassa 2 klustereiden määrä on suhteutettu sanoma- ja aikakauslehdissä julkaistujen merkkien määrään. Kansalliskirjaston digitointiprojektissa aineisto on tallennettu sivuttain, mutta sivujen ja kirjasimien koot muuttuivat voimakkaasti.[27] Jos lehtiä tarkastellaan informaation määrän suhteen, on selvää, että niiden ”vetoisuus” muuttui aikavälillä niin paljon, ettei sivumäärään suhteuttaminen tunnu mielekkäältä. Kuva 2 osoittaa, että vaikka toistettujen tekstien määrä kasvoi lähes eksponentiaalisesti 1800-luvun loppua ja 1900-luvun alkua kohden, merkkimäärään suhteutettuna kasvu oli tasaisempaa. Tämä viittaa siihen, että tekstien uudelleenkäyttö oli olennainen osa lehdistöä koko aikavälin ajan. Koska BLAST tunnistaa samuuksia OCR:n tason vaihteluista huolimatta ja koska sivunvaihdoista johtuvaa tekstien jakautumista useaan klusteriin on koko aikavälillä, kuvan 2 antamaa kuvaa voi pitää riittävän luotettavana. Selvää on toki, että useaan klusteriin jakaantuneita tekstejä on enemmän periodin loppupäässä, jossa lehdistön volyymi kasvoi.

Löytämämme klusteriaineisto sisältää kaikenlaisen toisteisuuden, myös erilaisten tiedotteiden, mainosten ja hinnastojen kaltaisten toistuvien tekstien kautta syntyvän samuuden. Toisteisuuden määrä sanomalehdistössä on kaiken kaikkiaan valtava, ja sen havaitseminen ja näkyväksi tekeminen tuo lehdistöhistoriaan ja suomalaisen julkaisukulttuurin tarkasteluun uuden ulottuvuuden, jonka merkitykset ovat moninaiset. Tekstien kierrättämisen kautta voidaan tarkastella muun muassa sitä, miten tieto on liikkunut paikasta toiseen, miten eri paikkakunnilla ilmestyneet lehdet

26. Tämän lisäksi hakutulokset on mahdollista tallentaa TSV-muodossa (*tab-separated values*), jolloin niihin voi soveltaa myös muita digitaalisten ihmistieteiden menetelmiä, kuten nimettyjen entiteettien tunnistamista (*named-entity recognition, NER*), paikkatietoanalyysia, aiheen mallinnusta (*topic modeling*) tai verkostoanalyysia.
27. Sanomalehdistön muutoksia voi tutkia Eetu Mäkelän kehittämällä työkalulla Finnish newspapers materiality explorer, https://newspaper-materiality-comhis.rahtiapp.fi/ (6.3.2019).

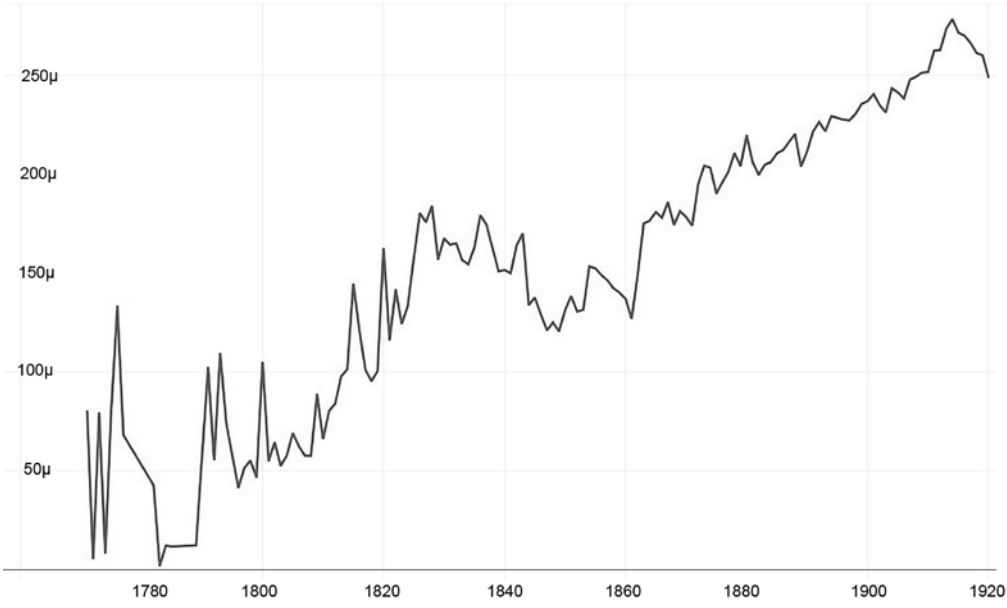


Kuva 1. Tekstien uudelleenkäytön klusterit suomalaisessa lehdistössä 1771–1920 (x = vuosi, y= klustereiden absoluuttinen määrä). Lähde: comhis.fi/clusters.

ovat seuranneet toistensa uutisointia ja millaisen kommunikaatioverkoston lehdistö on Suomen alueella muodostanut. COMHIS-tietokannassa on karttatyökalu, joka mahdollistaa klustereiden maantieteellisen laajuuden ja liikkeen visualisoinnin.[28] Työkalun avulla on mahdollista nähdä, miten sanoma- ja aikakauslehdet jakoivat toistensa sisältöä, ja informaation kulkua voi tarkastella vuositasolla tai pidemmän aikavälin kuluessa. Tarkastelun voi rajata myös kielen mukaan. Kuvassa 3 on kuvattu esimerkinomaisesti lehtien muodostamaa verkostoa vuonna 1877 siten kuin se hahmottuu tuona vuonna liikkeelle lähteneiden 59 581 suomen- ja ruotsinkielisen klusterin kautta (karttatoiminnossa pohjana on Googlen nykykartta). Jos tarkastellaan klustereiden ensimmäisiä julkaisupaikkoja, kolme vahvinta informaation kierrätyskeskusta olivat Helsinki (69 %), Turku (14 %) ja Viipuri (6 %). Toki lehtien välillä oli myös liikettä, joka ohitti suurimmat keskukset ja jota olisi tulevaisuudessa syytä tutkia tarkemmin. Koko tarkastelujaksolla Helsinki on selvästi suurin klustereiden lähtöpaikkakunta: noin 5,9 miljoonaa klusteria on lähtöisin helsinkiläisistä lehdistä. Seuraavaksi suurimpia klustereiden ensimmäisiä julkaisupaikkoja ovat Turku (noin 1,8 miljoonaa klusteria) sekä Viipuri (noin 1,2 miljoonaa klusteria). Esimerkiksi Tampereen selvästi pienempi painoarvo (noin 760 000 klusteria) on ymmärrettävää, sillä paikkakunnan ensimmäinen sanomalehti *Tampereen Sanomat* alkoi ilmestyä vuonna 1866.[29]

Aiemmassa tutkimuksessa on kiinnitetty huomiota kansainvälisen uutismateriaalin hyödyntämiseen ja suomalaislehtien erilaisiin uutiskanaviin. Kuvan 3 yhteydet näyttäytyvät Suomen "sisäisinä" siksi, että BLAST tunnistaa uudelleenkäyttöä vain korpuksen sisältä. Tosiasiassa ulkomaanuutisia saksittiin eli kopioitiin muualla julkaistuista lehdistä, ja ulkomaiset lehdet säilyivät tärkeänä uutislähteenä myös vuosisadan

28. Cluster Spread, http://comhis.fi/clusters/analysis/cluster_spread (6.3.2019).
29. Tulokset klusterien lähtöpaikoista ovat ylipäänsä samansuuntaisia aiemman tutkimuksen kanssa. Sanomalehtien lukumääristä ja niiden määrän vaihtelusta eri kaupungeissa ks. esim. Lars Landgren, Kieli ja aate. Politisoituva sanomalehdistö 1860–1889. Teoksessa Päiviö Tommila (toim.) *Suomen lehdistön historia* 1. S*anomalehdistön vaiheet vuoteen 1905*. Kustannuskiila 1988, 284–285, 389–390 ja passim.

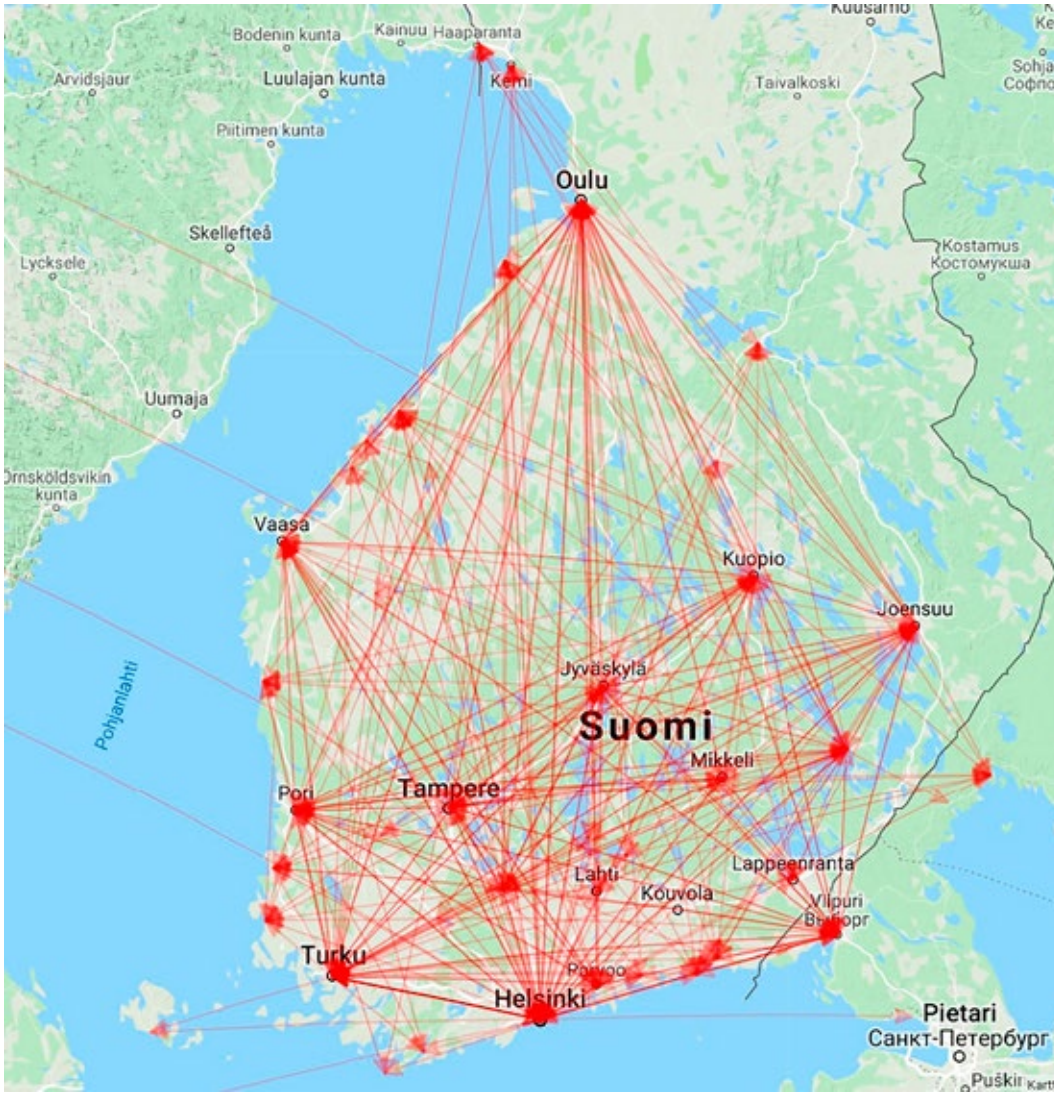


Kuva 2. Tekstien uudelleenkäytön klusterit suhteutettuna lehdistön merkkimäärään 1771–1920 (x = vuosi, y = klustereiden suhteellinen määrä). Lähde: comhis.fi/clusters.

jälkipuoliskolla teknologian kehityksestä huolimatta.[30] COMHIS-tietokanta näyttää ainoastaan Kansalliskirjaston materiaalissa ilmenevän toiston, mutta olemme havainneet osan kopioiduista uutisista kuuluvan osaksi laajempia kansainvälisiä toistoketjuja: tieto liikkui Euroopasta Suomeen esimerkiksi terveyttä uhkaavissa tai sotiin liittyvissä aiheissa.[31] Niitäkin kauemmas ulottuvaa kierrättämistä tutkitaan tällä hetkellä Ryan Cordellin johtamassa laajassa kansainvälisessä hankkeessa *Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840–1914*, jossa etsitään globaalien uutisvirtojen toistoketjuja ristiinlouhimalla eri maiden sanomalehtitietokantoja.[32]

Tiedon ja uutisten kopioiminen on ollut kansainvälisesti hyvin vahva, hyväksytty ja tavanomainen tapa koota uutistiedotteita ja säännöllisiä asiajulkaisuja.[33] Tekstien kopiointi synnytti ketjuja, jotka saattoivat myös haarautua ja muuntua, kun toimittajat kommentoivat tai lyhensivät tekstejä. Toistoketjut jatkuivat kielirajojen yli. Olemme huomanneet, ettei ollut lainkaan tavatonta, jos ruotsinkielinen uutinen julkaistiin suomeksi lyhennettynä ja joskus hyvinkin tiivistetysti käännettynä. Ulkomaanuutisten kohdalla tämä on helppo ymmärtää: ruotsinkieliset lehdet saattoivat nopeuttaa ulkomaiden tapahtumista raportointia kopioimalla laajan, kuvailevan tekstin Ruotsin lehdistä, mutta suomenkielisen lehden toimittajalta jokaisen vierasmaalaisen uutisen toisto vaati kääntämistä eli vaivannäköä. Tällöin tarkoituksenmukainen ajankäyttö usein lyhensi uutisointia suomeksi.[34]

30. Osmo Apunen, *Hallituksen sanansaattaja. Virallinen lehti – Officiella tidningen 1819–1969*. Valtion painatuskeskus 1970, 30; Tommila 1988, 110–111, 389.
31. Hannu Salmi, Asko Nivala, Heli Rantala, Reetta Sippola, Aleksi Vesanto & Filip Ginter, Återanvändningen av text i den finska tidningspressen 1771–1853. *Historisk Tidskrift för Finland* 1 (2018), 71–73.
32. Oceanic Exchanges Project Team, 2017, *Oceanic Exchanges. Tracing Global Information Networks In Historical Newspaper Repositories, 1840–1914*, https://osf.io/wa94s/ (27.6.2018).
33. Jonathan Silberstein-Loeb, Exclusivity and Cooperation in the Supply of News. The Example of the Associated Press, 1893–1945. *Journal of Policy History* 24:3 (2012), 467–468.
34. Victoria-laivan uppoamista Ontario-järvellä käsitellyt uutinen on esimerkki eripituisista uutisista (käännöksistä): mahdollisesti sähkeen perusteella oli laadittu yksityiskohtainen ruotsinkielinen uutinen, kun taas suomeksi julkaistiin parin



■ Kuva 3. Vuonna 1877 liikkeelle lähteneet klusterit. Visualisoinnissa on käytetty googlen karttapalvelua, ja siksi se näyttää nykyiset rajat ja paikannimet. Lähde: comhis.fi/clusters.

Tiedon liikettä vauhdittivat 1800-luvun viime vuosikymmeninä monet teknologiset muutokset. Rautatierata keisarikunnan pääkaupungista Pietarista Suomeen avattiin vuonna 1870, mikä nopeutti postin kuten sanomalehtien kuljetusta ja samalla uutisten leviämistä. Aiemmin tutkittujen viittausten perusteella sanomalehdet olivat Suomessa edelleen 1860- ja 1870-luvulla toisten lehtien tärkein uutiskanava. Ratatyöt jatkuivat Suomessa kiivaina seuraavat vuosikymmenet samalla tietoliikennettä parantaen. Samoin painotekniikka edistyi, ja sanomalehtien määrä alkoi kasvaa nopeasti. Tekstien liikettä tukivat ja lisäsivät samanaikaiset yhteiskunnalliset muutokset, erityisesti suotuisa taloudellinen kehitys,

lauseen tiedotus. Sama lyhyt sähke oli ensin julkaistu myös ruotsiksi. *Uusi Suometar* 27.5.1881, 3; *Åbo Underrättelser* 2.6.1881, 3. Ks. myös *Ilmarinen* 4.6.1881, 2 sekä Ruotsin sanomalehdet, https://tidningar.kb.se/ (6.3.2019). Alempana käsitelty uutinen Suomen ensimmäisestä naisinsinööristä kertoo samaa kotimaisella esimerkillä: kielenvaihto lyhensi uutista. Kielistä ja toimitustyöstä ks. ja vrt. Tommila 1988, 205–207.

joka kasvatti ilmoitusten ja mainonnan määrää.[35] COMHIS-tietokanta tarjoaa mahdollisuuden tarkastella myös mainosten volyymia, leviämistä ja kierrätystä Suomessa. Esimerkiksi sanalla ”alennusmyynti” tietokannasta löytyy 99 klusteria. Näistä laajin oli Arvi A. Kariston kirja-alennusmyyntimainos elokuussa 1917, joka julkaistiin 27 kertaa 16 paikkakunnalla.[36]

Näihin samanaikaisiin, rinnakkaisiin muutoksiin lukeutui myös sähköinen lennätin tai tarkemmin sen käytön tavanomaistuminen. 1700-luvun innovaatio optinen lennätin ei juuri kierrättänyt tekstejä. Sähköinen lennätinlinja Suomeen rakennettiin Pietarin suunnasta vuonna 1855. ”Sähkölanka” herätti huomiota uuden aikakauden ja yhteyksien symbolina, mutta se toimi käytännössä aluksi etupäässä sotilaskäytössä. Suomessa sähköinen lennätin vaikutti lehdistöön välillisesti, sillä tiedonkulku Pietariin ja Ruotsiin, ja samalla näiden lehdistä edelleen Suomeen, nopeutui. Sähköisen lennättimen vaikutus uutisvälityksessä alkoi näkyä sanomalehdissä 1870-luvun lopulla. Sähkösanomiin viitattiin esimerkiksi silloin, kun tieto J. L. Runebergin kuolemasta vuonna 1877 saavutti sanomalehdet: ”Tämä sanoma, joka sähkön woimalla on lentänyt ympäri Suomenmaan, on joka paikassa herättäwä surua ja kaipausta.”[37] Sähkösanomat pysyivät kuitenkin Suomessa edelleen kalliina ja niiden käyttöön liittyi rajoituksia. Ensimmäinen kotimainen uutistoimisto, Suomen Sähkösanomatoimisto, joka oli STT:n edeltäjä, aloitti vuonna 1887 tarkoituksenaan palvella uutisin nimenomaan maaseudulle syntyneitä sanomalehtiä. Sähköisen lennättimen käyttö alkoi siten yleistyä samanaikaisesti puhelimen kanssa, johon uudenaikaisimmat lehtien toimitukset olivat tarttuneet jo vuonna 1882.[38] Nämä tiedonkulun ja -siirron nopeutumisen ja halpenemisen monet osatekijät näkyvät tietokannassa toistoketjujen määrän merkittävänä ja jatkuvana kasvuna 1800-luvun lopulla ja 1900-luvun alussa.

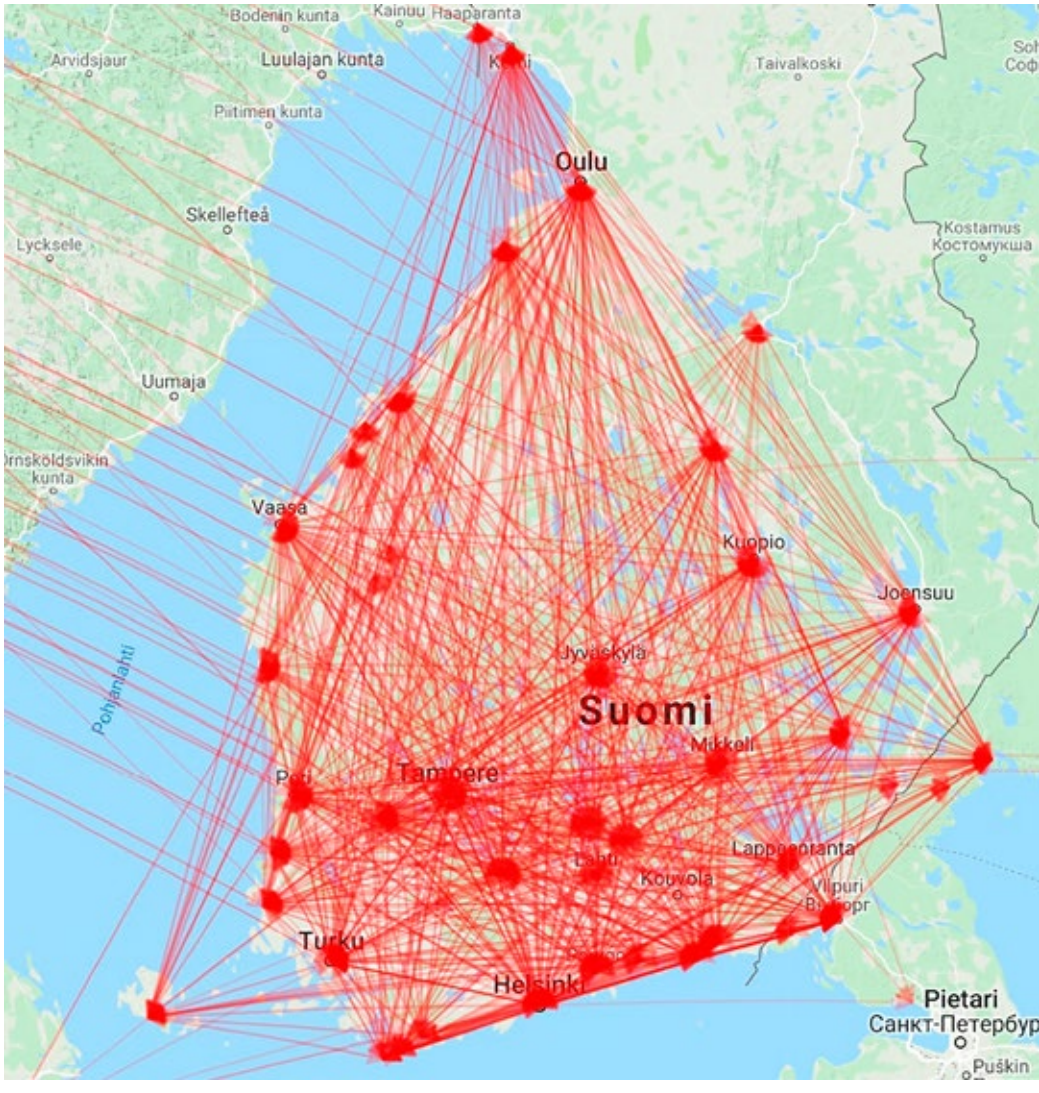
Kuvassa 4 nähdään uudelleenkäytön verkosto vuonna 1897, 20 vuotta kuvan 3 tilannetta myöhemmin. Uudelleenkäytön klustereita oli 240 933 kappaletta vuonna 1897, nelinkertainen määrä vuoteen 1877 nähden. Kartassa on myös säikeitä, jotka lähtevät rajauksen ulkopuolelle. Aineistossa ovat mukana Pohjois-Amerikassa julkaistut aikakauslehdet ja, kuten klusteriaineistosta voi päätellä, moni 1890-luvun uutisketju on jatkunut siirtolaislehdistön mukana Atlantin toisella puolen.

Tulevina vuosina lehdistön kasvu ja volyymi jatkuivat, mitä myös uudelleenkäyttö kuvastaa. Vuonna 1907 klustereita lähti liikkeelle 492 157 kappaletta ja vuonna 1917 lukumäärä oli 530 843 kappaletta. Kuten todettu, luvut eivät viittaa suoraan julkaistujen tekstien määriin, sillä tekstit voivat olla jakaantuneita useampaan klusteriin. Ne ovat kuitenkin suuntaa-antavia ja kertovat uutisten maantieteellisistä liikkeistä ja tihenevästä tiedonvälityksestä.

Viraalisuuden jäljillä

Viraalisuuden käsite on tullut tutuksi 2000-luvun mediakulttuurista, jossa se viittaa ”kuvan, videon tai tiedon nopeaan ja laajaan leviämiseen internetissä käyttäjältä toiselle”.[39] Voi kuitenkin väittää, että viraalisuuden mahdollisuusehto, viestien nopea leviäminen, toteutui jo sanomalehdistön laajenemisen yhteydessä 1800-luvulla.[40] Suomessa viraalisuudesta voi kuitenkin puhua vasta 1800-luvun lopussa, jolloin median kapasiteetti voimakkaasti laajeni. Esimerkkinä voidaan ajatella kirjettä, jonka Suomalaisen Kirjallisuuden Seura laittoi liikkeelle elokuussa 1864 Porthanin patsaan paljastamisesta. Kiertokirje julkaistiin lopulta 18 sanomalehdessä, siis kaikissa niissä lehdissä, jotka ylipäätään ilmestyivät elokuussa 1864. SKS:n kirje sai maksimaalisen näkyvyyden, mutta kovin viraalina sitä ei voi pitää, jos kriteerinä on leviämisen absoluuttinen volyymi. Viraalisuuden ominaisuuksiin kuuluu kuitenkin myös

35. Marja-Liisa Suomalainen, Sanomalehtien uutiskanavista 1860- ja 1870-luvuilla. *Lehdistöhistoriallisia tutkimuksia* 1. Suomen sanomalehdistö historia -projektin julkaisuja 1979, 103–140; Landgren 1988, 267–420, 280–283, ja passim.
36. Klusteri 15082100, http://comhis.fi/clusters/?f%5Bcluster_id%5D%5B%5D=15082100 (6.3.2019).
37. *Länsi-Suomi* 12.5.1877, 1; Suomalainen 1979, 105.
38. Einar Risberg, *Suomen lennätinlaitoksen historia* 1855–1955. Posti- ja lennätinhallitus 1959, 22–25, 37, 46–49; Tapani Kaskinen, *Lennätin ja radio. Tietoliikenne Suomessa 1860–1939*. Suomen sanomalehdistön historia -projekti 1978, 76–93; Terhi Rantanen, *”STT:n uutisia” sadan vuoden varrelta*. Weilin & Göös 1987, 20–23; Kari Immonen, *Sillat sielujen ja ihmismietteen. Suomalaisen puhelimen kulttuurihistoriaa keskusneideistä tekstiviesteihin.* Edita 2002, erit. 64–66.
39. Ks. lähemmin *Oxford Dictionary of English*. 3. verkkopainos. Oxford University Press 2015, DOI: 10.1093/acref/9780199571123.001.0001 (6.3.2019).
40. Ks. lähemmin Salmi 2018, 71–79.



■ Kuva 4. Vuonna 1897 liikkeelle lähteneet klusterit. Visualisoinnissa on käytetty googlen karttapalvelua. Lähde: comhis.fi/clusters.

nopeus: tämän tekstin leviäminen kesti 15 päivää, vaikkakin useimmat ilmestyivät viiden päivän kuluessa.[41]

Lehdistön volyymi kasvoi merkittävästi 1860-luvun jälkeen. Tarkastelujaksomme lopussa vuonna 1920 sanomalehtiä ilmestyi 136 nimikettä. Saman vuoden aikakauslehtiä on Kansalliskirjaston digikokoelmassa 412 nimikettä. Rakentamassamme tekstien uudelleenkäytön tietokannassa on mahdollisuus selata klustereita sekä toistomäärän (*count*) että viraalisuuden (*virality score*) avulla. Korkein viraalisuusluku on ilmoituksella, jossa suomalaiset tupakkatehtailijat vetosivat kotimaisten savukkeiden puolesta taistelussa ”amerikkalaisen tupakkatrustin maassamme juurtumista vastaan”.[42] Ilmoitus julkaistiin 75 kertaa, 45 eri sanomalehdessä 26 paikkakunnalla, ja se oli jatkoa julkisuudessa käydylle kamppailulle

41. Tapauksesta tarkemmin ks. Salmi 2018, 75–76.
42. Ks. esim. *Kotimainen työ* 3/1916, 43, sekä samasta ”Taistelu tupakkatrustia vastaan,” s. 41. Ks. lähemmin klusteri 11592519, http://comhis.fi/clusters/?f%5Bcluster_id%5D%5B%5D=11592519 (6.3.2019).

siitä, miten suhtautua ulkomaisen jättiyhtiön tuotteisiin. Kyse ei toisin sanoen ollut lehdistön sisällä syntyneestä ilmiöstä, vaan tarkoitus oli saada tärkeänä pidetty asia leviämään nopeasti mahdollisimman laajalle lehdistön kautta. Ilmoitus julkaistiin ensin päivälehdissä *Keski-Savo* (Savonlinna), *Tampereen Sanomat* ja *Suupohjan Kaiku* (Kristiinankaupunki) 14. maaliskuuta 1916 ja sen jälkeen nopeana sarjana maaliskuun loppuun 1916 saakka.[43]

Monet korkean viraalisuuden tekstit olivat ilmoituksia tai tiedotteita. Kaiken kaikkiaan klustereita, joissa viraalisuusindeksi on yli 50, on tietokannassa 81 kappaletta. Näistä selvästi levinnein journalistinen teksti on *Käkisalmen Suomalaisessa* ja *Kajaanin Lehdessä* Aleksis Kiven päivän aattona 9. lokakuuta 1911 ilmestynyt Yrjö Koskelaisen (1885–1951) kirjoitus suomalaisesta kirjallisuudesta.[44] Tässäkin tapauksessa teksti on levitetty tietoisesti lehdille mahdollisimman laajan julkisuuden saavuttamiseksi. Samalla tekstin levinneisyys kertoo, miten sanomalehdistö osallistui aktiivisesti Aleksis Kiven maineen ja suomalaisen kulttuurin rakentamiseen. Nuorsuomalaisiin kuulunut Koskelainen oli samaan aikaan *Valvoja-* ja *Aika*-lehtien avustaja, joka seuraavana vuonna tuli *Uuden Auran* toimittajaksi.[45] Koskelaisen kirjoitus julkaistiin 64 kertaa lokakuun 9. ja 14. päivän välisenä aikana ja vielä kerran vuotta myöhemmin. Koska julkaisutiheys oli näin nopea, on selvää, että tekstin leviämisessä ei ollut kyse kopioinnista lehdestä toiseen, mutta tekstin tehokasta kierrättämistä se oli.

Tietokanta osoittaa, miten sanomalehdistö oli ilmoitusten, mainosten, tiedotteiden ja propagandankin kaikupohja. Lehdistö tarjosi laajan julkisuuden, ja sen kautta oli mahdollista saavuttaa nopeasti suuri yleisö eri puolilta Suomea. Toki oli myös tekstejä, jotka levisivät lehdestä toiseen, jolloin rytmi oli hitaampi. Tällainen oli esimerkiksi vuonna 1911 levinnyt kirjoitus ”Talvihuomautuksia maanmiehille”, joka oli kopioitu ruotsalaisesta lehdestä ja jonka tarkoituksena oli tarjota ohjeita talviajan arkielämään.[46] Viraalisuusarvo on vain 2.07, mikä selittyy uutisen hitaammalla etenemisellä, koska se kopioitiin nimenomaan lehdestä toiseen. Teksti ilmestyi Suomessa aluksi *Suomalainen Kansa* -lehdessä 1. helmikuuta 1911, ja *Etelä-Suomi* -lehdessä vasta kolme päivää myöhemmin. Lopulta Ruotsista peräisin ollut kirjoitus julkaistiin 30 kertaa 21 päivän kuluessa 19 paikkakunnalla.

Lehdistön viraalisuudesta kertoo myös vuonna 1905 julkaistu koulutusuutinen. Toukokuun viimeisenä päivänä vuonna 1905 sanomalehti *Helsingfors-Posten* nosti neiti Jenny Markelinin esiin vastavalmistuneiden insinöörien joukosta kertomalla, että hän oli ”Finlands första kvinnliga ingenjör”. Useampien naisarkkitehtien jälkeen Markelin oli ensimmäinen insinööriosastolla tutkinnon suorittanut nainen. Seuraavana päivänä pääkaupunkiseudun suurimmat ruotsin- ja suomenkieliset sanomalehdet toistivat uutisen Suomen ensimmäisestä naisinsinööristä lähes sanasta sanaan. Pian uutinen kopioitiin julkaistavaksi eri alueiden lehdissä ympäri valtakuntaa Sortavalaa (*Laatokka*-lehti) ja Kuopiota myöten. Jälkimmäisessä kaupungissa ilmestynyt lehti *Pohjois-Savo* toisti uutisen suomenkielisen version ketjun viimeisenä 5. kesäkuuta 1905. Ylipäänsä myöhäisin uutisen toisto oli *Kotka Nyheter* -lehdessä 10. kesäkuuta 1905.[47]

Turussa *Uusi Aura* toisti 1. kesäkuuta 1905 vain osan uutisesta. Tapaus tuo esiin myös tietokannan rajaukset, sillä alle 300 merkin uudelleenkäyttöä ei voitu tunnistaa. Kansalliskirjaston digitaalinen kokoelma sisältääkin useita lyhempiä, osittaisia toistoja, jotka eivät ole päätyneet tietokantaamme. Kun aineistostamme löytyy yhteensä 16 uutisen toistoa kahdella kielellä, Kansalliskirjaston digitoiduista sanomalehdistä ilmenee lisäksi muita uutisen toisintoja noin kymmenen samalta ajanjaksolta. Kaiken kaikkiaan tieto Suomen ensimmäisestä naisinsinööristä saavutti

43. *Keski-Savo* 14.3.1916, 2; *Tampereen Sanomat* 14.3.1916, 1; *Suupohjan Kaiku* 14.3.1916, 1. Samaan toistoketjuun kuuluva *Kotimainen työ* 3/1916, 43 näyttää tietokannassa varhaisimmalta ilmoitukselta klusterissa 11592519, mutta kyseessä on jälkikäteen tuotettu, virheellinen metatieto (1.3.), sillä ko. numero ilmestyi jonakin päivänä maaliskuussa mutta tuskin kuun ensimmäisenä päivänä. Tämän huomioiminen vain kasvattaa ilmoituksen viraalisuutta. Ks. aikakauslehtien epätarkoista metatiedoista laajemmin edellä.
44. Ks. lähemmin klusteri 13611711, http://comhis.fi/clusters/?f%5Bcluster_id%5D%5B%5D=13611711 (6.3.2019).
45. Jukka Muiluvuori & Yrjö Koskelainen, *Kansallisbiografia-verkkojulkaisu*. Studia Biographica 4. Suomalaisen Kirjallisuuden Seura 1997– (15.6.2018).
46. Ks. lähemmin klusteri 10634989, http://comhis.fi/clusters/?f%5Bcluster_id%5D%5B%5D=10634989 (6.3.2019).
47. *Helsingfors-Posten* 31.5.1905, 3.

suuren osan suuriruhtinaskuntaa kesäkuun ensimmäisellä viikolla vuonna 1905. Sanomalehtien jälkeen uutista toistivat muokaten jotkin aikakauslehdet kuten *Nutid*, joka oli Naisasialiitto Unionin äänenkannattaja, *Palvelijatarlehti* ja *Suomen Teollisuuslehti*.[48] Tämä esimerkki muistuttaa myös siitä, että tietokannan toistomateriaalia kannattaa tarkastella rinnan Kansalliskirjaston alkuperäisen digitoidun aineiston kanssa.[49]

Sanoma- ja aikakauslehdistö kierrätti tekstejä 1800-luvun loppua kohti entistä nopeammin. Kierrätys täytti viraalisuuden tunnusmerkit, ja uutisia, ilmoituksia ja tiedonantoja levitettiin nopeassa tahdissa maantieteellisesti laajalla alueella. Lehdistölle syötettiin ja tarjottiin tekstejä kierrätettäväksi, mutta samalla lehdet kopioivat tekstejä myös toisiltaan. Kopioinnin kautta ja laatimamme tietokannan avulla voi analysoida, miten ilmiöt voimistuivat 1800-luvun lopun ja 1900-luvun alun Suomessa.

Pitkän aikavälin uudelleenkäyttö

Eräänlaisena viraalisuuden vastakohtana suomalaisista lehdistä löytyy toistoa ja lainaamista, joka ilmenee hyvin pitkällä aikavälillä, vuosikymmenten tai jopa yli sadan vuoden aikajänteellä. On vaikea arvioida, onko tämä pitkän aikavälin toisteisuus leimallista erityisesti suomalaiselle lehdistölle, sillä kansainvälisiä vertailukohteita ei juuri ole.[50] Varhaisimpien suomalaisten lehtien kohdalla synkronista lainaamista lehdestä toiseen ei juuri löydy rinnakkaisten julkaisujen vähyyden takia. Tämä kierrätyksen poissaolo antaa silti vääristyneen kuvan 1800-luvun alun lehdistä, sillä kuten jo todettu, lehdet lainasivat materiaalia ulkomaisista lehdistä. Näissä tapauksissa kopioimiseen sisältyi myös tekstien kääntämistä kieleltä toiselle. 1800- ja 1900-lukujen vaihteen lehdet puolestaan julkaisivat uudelleen 1700-luvun lopun ja 1800-luvun alun varhaisten suomalaislehtien materiaalia. Kasvava lehdistö näyttää ammentaneen materiaalia varhaisista lehdistä julkaisemalla joko joitain osia tai jopa kokonaisia kirjoituksia 1700- ja 1800-lukujen taitteen sanomalehdistä. Osa tästä uudelleenkäytöstä on yksittäistapauksia, jolloin vanha teksti on uudelleenjulkaistu vain kerran. Joissain tapauksissa vanha teksti on sen sijaan muuttunut uudessa julkaisuympäristössään paljon toistetuksi.

Uudelleenkäytön kokonaisvolyymissa pitkän aikavälin toisteisuus ei ole erityisen voimakas piirre. Kun aineiston sisältämien uudelleenkäyttöklustereiden määrä lasketaan miljoonissa, puhutaan pitkän aikavälin toistossa tuhansista klustereista. Tässä artikkelissa tarkoitamme pitkällä aikavälillä uudelleenkäyttöä, jossa tekstin ensimmäisen ilmestymisen ja viimeisen toistotapauksen välinen aikajänne (tietokannassa *span*) lasketaan vähintään kymmenissä vuosissa. 85 prosenttia klustereista sisältää uudelleenkäyttöä, joka tapahtui vuoden sisällä. Vähintään kymmenen vuoden aikajänteen sisältäviä klustereita tietokannasta löytyy lähes 260 000 ja 40 vuoden viiveen tapauksia vielä lähes 12 000. Sen sijaan 50 vuoden aikavälin klustereita löytyy enää 5 888, mikä sekin on sinänsä huomattava määrä. Klustereita, joiden *span*-arvo on sata vuotta, löytyy tietokannasta 289 kappaletta. *Span*-arvon lisäksi pitkän toiston tapauksissa merkittävää on tekstin ensimmäisen julkaisemisen ja uudelleenjulkaisun välinen etäisyys (tietokannassa *gap*), joka tarkoittaa julkaisuajankohtien välistä taukoa. Pisimmillään tämä ”hiljaisuus” on tietokannan aineistossa yli 145 vuotta. Luultavasti tätäkin pidempiä toistoklustereita olisi löydettävissä, jos käytössä oleva aineisto ulottuisi vuotta 1920 pidemmälle.

Pisimmän *span*- tai *gap*-arvon toistoketjut ovat käytännössä tapauksia, joissa 1900-luvun alun lehdet ovat hyödyntäneet vanhimpien 1700-luvun lopun lehtien materiaalia. Esimerkki erittäin pitkän aikavälin toistosta on klusteri (11 250 221), jonka *span* on 145 vuotta ja toistomäärä (*count*) 21. Kyse on *Suomenkieliset Tieto-Sanomat* -lehden näytenumerossa vuonna 1775 ilmestyneestä lehden aloitustekstistä ja näin ollen aivan ensimmäisestä suomen kielellä julkaistusta sanomalehtitekstistä.[51] Klusterin mukaan tekstin ensimmäinen uudelleenjulkaisija oli *Suometar*

48. *Nutid* 6–7/1905, 237–238; *Palvelijatarlehti* 3–4/1905, 45; *Suomen Teollisuuslehti* 12/1905, II.
49. Petri Paju, Ensimmäiset naiset insinöörien ja arkkitehtien yhdistyksissä. *Tekniikan Waiheita* 36:1 (2018), 5–24.
50. Yhdysvaltalaisessa materiaalissa on havaittu lainaamisen tapahtuneen joissain tapauksissa useamman vuoden aikajänteellä. Suomen kaltaista pitkää toisteisuutta ei ole havaittu, sillä käytetty korpus oli vuosilta 1830–1860. Smith, Cordell & Maddock Dillon 2013, 93.
51. Lehden näytenumero ilmestyi syyskuussa 1775, ja vuoden 1776 aikana lehti ilmestyi 24 numeron verran. Lehdestä tar-

vuonna 1858, jonka jälkeen sama tekstinpätkä julkaistiin yhden tai useamman kerran vuosina 1866, 1895, 1904, 1905, 1909, 1912, 1913, 1915, 1918 ja 1920 yhteensä 18 eri lehdessä. Yli 130 vuoden aikajänteellä julkaistuissa uudelleenkäyttötapauksissa hallitsevana ovat juuri *Tieto-Sanomat*-lehdestä lainatut tekstit. Myös 1700-luvun *Tidningar*-lehden materiaalia julkaistiin uudelleen pitkällä aikavälillä.[52] Näiden varhaisten lehtien materiaalin uudelleenjulkaisemisen motiiviksi lienee osittain riittänyt se, että kyseessä olivat ensimmäiset Suomen alueella ilmestyneet sanomalehtijulkaisut. Kierrättämällä näiden lehtien aineistoa 1800- ja 1900-lukujen vaihteen lehdistö on luonut ja ylläpitänyt tietoisuutta suomalaisen sanomalehtijulkaisemisen historiasta. Uudelleenjulkaisuissa on havaittavissa myös tiettyjen vuosipäivien muistamista.[53]

Pitkän aikavälin klustereiden joukossa on tapauksia, jossa vanha lehtiteksti on toistettu myöhemmin vain kerran. Esimerkiksi vuonna 1910 Suomen purjehdusseuran julkaisema *Frisk Bris* -aikakauslehti julkaisi meren tulvimista käsittelevän tekstin, joka ilmestyi alun perin *Tidningar*-lehdessä vuonna 1771. Julkaisuajankohtien välissä on 139 vuoden hiljaisuus. *Finsk musikrevy* -aikakauslehti on puolestaan julkaissut katkelmia *Tidningar*-lehden musiikkiaiheisista kirjoituksista useammassa yhteydessä. Vuonna 1906 se lainasi osia turkulaislehdessä vuonna 1773 ilmestyneestä konserttiuutisesta. Näissä tapauksissa uudelleenjulkaiseminen liittyy selkeästi kyseisen aikakauslehden aihepiiriin ja kyse on yksittäisestä lainaamisesta. Sen sijaan esimerkiksi Elias Lönnrotin *Mehiläisessä* vuonna 1837 julkaistu paloviinan käytön ja juoppouden vaaroista muistuttava teksti eli myöhemmässä lehdistössä pitkään useiden toistojen ansiosta. Juttua toistettiin vuosien 1889 ja 1902 välisenä aikana, raittiusliikkeen kultakautena, useammassa syklissä yli 30 kertaa. Tässä tapauksessa tekstin runsaaseen toistamiseen vaikutti aiheen lisäksi varmasti myös se, että kirjoitus oli alun perin ilmestynyt Lönnrotin lehdessä.

Osa lehdistössä pitkään kiertäneistä teksteistä on erilaisia anekdootteja tai pikku-uutisia, jotka on selvästi lainattu suomalaislehtiin ulkomaisista lehdistä ja joiden kohdalla kysymys alkuperästä tai tekijyydestä on moniulotteinen eikä välttämättä edes kovin mielekäs. Ryan Cordell on esittänyt yhdysvaltalaisen aineiston pohjalta erilaisten anonyymien pikkujuttujen kierrättämisen olleen yleinen tekstin uudelleenkäytön muoto 1800-luvun alkupuolen sanomalehtijulkisuudessa.[54] Ilmiö näkyy myös suomalaisessa aineistossa, vaikka sen tarkkaa painoarvoa on vaikea arvioida ilman klustereiden jakamista eri tekstilajeihin. Joissain tapauksissa vanha teksti on uudelleenkäytön myötä saanut varsin laajan ja nopean levityksen. Vuonna 1851 *Suometar* julkaisi lyhyen kirjoituksen Askolassa toimivasta räätälistä, joka oli erityisen nopea juoksija.[55] Teksti nousi uudelleen sanomalehtijulkisuuteen vuonna 1913, jolloin sen julkaisi ensimmäisenä *Uusi Aura* 16. heinäkuuta 1913. Tämän jälkeen teksti kiersi lehdissä tehokkaasti, ja heinä-elokuun aikana tarina nopeajalkaisesta räätälistä ilmestyi yhteensä 29 eri lehdessä ympäri Suomen. Teksti oli ilmestynyt ruotsinkielisenä *Borgå Tidning* -lehdessä jo 21. toukokuuta 1851, josta *Suometar* oli sen ilmeisesti kääntänyt, sillä lehti mainitsi lähteekseen Porvoon lehden. Koska BLAST ei pysty etsimään samuuksia eri kielten välillä, ei tätä ruotsinkielistä tekstiä löydy kyseisen tekstin klusterista vaan olemme identifioineet varhaisimman julkaistun version manuaalisesti Kansalliskirjaston lehtitietokannasta.

BLAST:in ansiosta näkyväksi tuleva pitkän aikavälin toisteisuus osoittaa, että lehdistön merkitys ei ollut vain uutuusarvossa ja nopeudessa vaan myös muistamisessa. Tätä puolta lehdistöstä olisi vaikea tai mahdotonta hahmottaa ilman konelukemisen ja laskennallisten menetelmien apua. Tekstien pitkä kiertäminen suomalaisessa lehdistöjulkisuudessa on esimerkki havainnosta, jonka algoritmi on mahdollistanut ja nostanut esiin.

kemmin, ks. esim. Jyrki Pietilä, *Kirjoitus, juttu, tekstielementti. Suomalainen sanomalehtijournalismi juttutyyppien kehityksen valossa printtimedian vuosina 1771–2000*. Jyväskylä studies in humanities 111. Jyväskylän yliopisto 2008, 105–108.

52. Aiheesta tarkemmin ks. Heli Rantala, "Porthanin lehti" ja otteista sen myöhemmästä elämästä. *Auraica. Scripta a Societate Porthan Edita* 9 (2019), 59–65, https://journal.fi/aur/article/view/78061 (6.3.2019).

53. Tästä tarkemmin, ks. Hannu Salmi, Heli Rantala, Aleksi Vesanto & Filip Ginter, The Long-Term Reuse of Text in the Finnish Press, 1771–1920. *Proceedings of the 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019.* (tulossa)

54. Cordell 2015, 423–424, 429–430.

55. *Suometar* 10.6.1851, 3–4.

Lopuksi

Tutkimuksemme osoittaa, että lainaaminen ja kierrättäminen ei ole liittynyt vain johonkin tiettyyn vaiheeseen suomalaisen lehdistön historiassa, vaan se luonnehtii koko tutkittua aikaväliä. Sitä mukaa kun lehdistön kokonaisvolyymi kasvoi 1800-luvun kuluessa, myös jaettujen tekstien määrä kasvoi. Erityinen piirre, jota käsittääksemme ei ole analysoitu aiemmin, on pitkän aikavälin toisto: lehdistö oli arkisto, josta voitiin ammentaa sisältöjä uusiin julkaisuihin, ja samalla siitä rakentui kulttuurisen muistin kanava. Lehdistössä ilmeni myös viraalista toisteisuutta, jossa sama teksti levisi verkostossa nopeasti, usein muutamien päivien tai viikkojen kuluessa. Koska jaettuja tekstejä on miljoonia, päätimme julkaista aineiston tietokantana, jotta tulokset voivat hyödyttää tutkimusta laajemmin.

Kierrätetty aineisto johtaa pohtimaan toimijuuden merkitystä. Lehdistö oli 1800-luvulla paitsi monien tekstilajien myös monenlaisten intressien foorumi. Ilmoitusten ja tiedonantojen suuri määrä korostaa materiaalin moniäänisyyttä ja jaettua toimijuutta: lehdistön kautta puhuivat yritykset ja yhteisöt, seurakunnat ja kaupungit, viranomaiset ja kansalaiset. Journalistisen kirjoittamisen ja mielipidevaikuttamisen merkitys kasvoi tutkitulla aikavälillä, mutta luonteeltaan lehdet olivat heterogeenisiä. Yhdysvaltain 1800-luvun sanomalehdistöä tutkinut Ryan Cordell on käyttänyt käsitettä ”network author” kuvaamaan yhteisöllistä ja jaettua tekijyyttä.[56] Myös Suomessa lehdistöä voisi tulkita verkostotekijyyden kautta: kirjoittajuuteen vaikuttivat toki toimittajat ja muut mielipidevaikuttajat, mutta yhtä lailla ne rakenteelliset tavat, joiden kautta tekstit syntyivät.[57] Tähän jaettuun toimijuuteen kuuluivat kierrätyksen painopisteet ja keskukset, tiedonkulun ja informaation saatavuuden reitit sekä monet uudelleenjulkaisemisen käytännöt.

Menetelmänä tekstien uudelleenkäytön tunnistus on hedelmällinen keino tutkia informaation liikkeitä ja reittejä. Käyttämällämme menetelmällä on kuitenkin myös rajoituksia, kuten se, että kääntämistä eli uudelleenkäyttöä kielirajan yli, ei ole voitu tunnistaa automaattisesti. Sama uutinen on voinut levitä sekä suomen- että ruotsinkielisenä, kuten esimerkki Suomen ensimmäisen naisinsinöörin valmistumisesta vuonna 1905 osoittaa. Tällä hetkellä optisen tekstintunnistuksen laatu vaikeuttaa merkittävästi automaattista käännösten löytämistä. Käännösten tunnistus olisi tärkeää myös sen ymmärtämiseksi, miten suomalainen lehdistö osallistui kansainväliseen uutisten ja tiedon vaihtoon. Suomi ei ollut vain vastaanottava alue, sillä uutisia levisi myös Suomesta muualle.[58] Jos olisi mahdollista laajemmin yhdistää digitoitujen sanomalehtien korpuksia, tulisivat näkyviin ne kansainväliset tiedonkulun verkostot, joiden osaksi suomalainen lehdistö syntyi.

FT **Heli Rantala** työskentelee tutkijatohtorina Turun yliopiston kulttuurihistorian oppiaineessa. **Sähköposti:** heli.rantala@utu.fi.

Hannu Salmi on akatemiaprofessori Turun yliopiston kulttuurihistorian oppiaineessa. **Sähköposti:** hannu.salmi@utu.fi.

Dosentti **Asko Nivala** työskentelee tutkijatohtorina Turun yliopiston tutkijakollegiumissa (TIAS). **Sähköposti:** asko.nivala@utu.fi.

Dosentti **Petri Paju** toimii tutkijana Turun yliopiston kulttuurihistorian oppiaineessa. **Sähköposti:** petri.paju@utu.fi.

FM **Reetta Sippola** on tohtorikoulutettavana Turun yliopiston kulttuurihistorian oppiaineessa. **Sähköposti:** raniem@utu.fi.

Aleksi Vesanto on valmistunut maisteriksi Turun yliopiston tulevaisuuden teknologioiden laitokselta. **Sähköposti:** aleksi.vesanto@utu.fi.

Dosentti **Filip Ginter** on kieli- ja puheteknologian apulaisprofessori Turun yliopiston tulevaisuuden teknologioiden laitoksella. **Sähköposti:** filip.ginter@utu.fi.

56. Cordell 2015, 417–445.
57. Toimituksellisista prosesseista ks. lähemmin Heidi Kurvinen, Toimittajat ja toimitukselliset prosessit mediatekstien takana, *Historiallinen Aikakauskirja* 3 (2018), 310–322.
58. Erinomainen esimerkki on Turun palon kansainvälinen uutisointi, joka syksyllä 1827 ja keväällä 1828 sai globaalit mittasuhteet. Ks. tarkemmin Hannu Salmi, Catastrophe, Emotions and Guilt. The Great Fire of Turku 1827. Teoksessa Deborah Simonton & Hannu Salmi (toim.) *Catastrophe, Gender and Urban Experience, 1648–1920.* Routledge 2017, 121–138.