Suomi24 : muodonantoa aineistolle

Lagus, Krista Hannele

2016-05

Lagus, K H, Ruckenstein, M S, Pantzar, M & Ylisiurua, M J 2016, Suomi24 : muodonantoa aineistolle. Valtiotieteellisen tiedekunnan julkaisuja, Nro 10, Helsingin yliopisto, Helsinki. < http://hdl.handle.net/10138/163190 >

http://hdl.handle.net/10138/163190

publishedVersion

10
2016

Krista Lagus & Mika Pantzar &
Minna Ruckenstein & Marjoriikka Ylisiurua

SUOMI24

Muodonantoa aineistolle

Valtiotieteellisen tiedekunnan julkaisuja

SUOMI24

MUODONANTOA AINEISTOLLE

Krista Lagus, Mika Pantzar,
Minna Ruckenstein, Marjoriikka Ylisiurua

Kuluttajatutkimuskeskus 2016:10
Valtiotieteellisen tiedekunnan julkaisuja, Helsingin yliopisto.

Helsinki 2016

Valtiotieteellisen tiedekunnan julkaisuja 10 (2016)
Kuluttajatutkimuskeskus

Kansi: Riikka Hyypiä & Hanna Sario

Tilastolliset analyysit: Pasi Karhu, StudioMind Oy

ISSN 2343-273X (painettu)
ISSN 2343-2748 (verkkojulkaisu)
ISBN 978-951-51-1065-7 (nid.)
ISBN 978-951-51-1066-4 (pdf)

Unigrafia, Helsinki 2016

ESIPUHE

Tässä raportissa luodaan katsaus Suomi24-keskustelufoorumin aineistoon, joka on ollut saatavilla Kielipankissa keväästä 2015. Raportissa kuvataan aineiston syntytapaa ja kontekstia, eli Suomi24-foorumia, aineiston määriä ja luonnetta, aineiston tallennusmuotoja, ja tapoja, joiden avulla tutkija voi aineistoon perehtyä. Lisäksi viritetään esimerkinomaisesti joitakin yhteiskuntatieteellisiä näkökulmia, joihin tutkimus voisi edetä.

Aineiston käyttäjille raportti toivottavasti tarjoaa hyödylliset perustiedot aineistosta ja sen ominaisuuksista, sekä tavan viitata tähän tutkimuksen resurssiin. Raportti tarjoaa taustan, jota vasten yksittäisten hakujen tuloksia voi suhteuttaa ja tulkita. Lisäksi aineiston käyttöesimerkkien ja erilaisten lähestymistapojen esittely voi tarjota virikkeitä aineiston hyödyntämiselle omassa tutkimuksessa.

Tätä työtä on tehty syksyllä 2015 osana Mika Pantzarin johtamaa ”Kansakunnan tunneaallot” -hanketta, jota on rahoittanut Teknologiateollisuus 100-v -rahasto. Hanke on osa laajempaa Citizen Mindscapes-tutkimuskollektiivia, joka osallistui Helsinki Challenge –tiedekilpailuun 2014–2015. Tässä käynnistettyä työtä jatketaan ja laajennetaan vuosina 2016–2018 ”Kansakunnan mielenliikkeet”-konsortiohankkeessa joka on osa Suomen Akatemian Digitaalisten ihmistieteiden ohjelmaa. Kiitämme kaikkia yksilöitä, hankkeita, ryhmiä, yhteisöjä, tahoja ja organisaatioita, jotka ovat tukeneet tätä tutkimusta. Keskustelu ja yhteistyö jatkukoon kaikkia osapuolia rikastuttaen eri yhteyksissä ja foorumeilla.

Helsingissä, tammikuussa 2016, raportin kirjoittajat

SISÄLLYS

1. MITÄ ON SUOMI24-AINEISTO?

Suomi24-aineisto sisältää Suomi24-keskustelufoorumin[1] keskusteluja vuodesta 2001. Foorumi ja aineisto ovat Aller Oy:n omistuksessa. Helsingin yliopiston Politiikan ja talouden tutkimuksen laitoksella toimiva Kuluttajatutkimuskeskus on huhti-toukokuussa 2015 avannut yhteistyössä Allerin, FIN-CLARIN:in, Helsingin Yliopiston Valtiotieteellisen tiedekunnan Menetelmäkeskuksen sekä CSC-Tieteen tietotekniikan keskuksen kanssa Suomi24-aineiston tutkimuskäyttöön. Yhteistyötä on tehty osana Citizen Mindscapes-tutkimuskollektiivin toimintaa.

Aineisto on tarkoitettu ei-kaupalliseen opetus- tai tutkimustarkoitukseen. Tavoitteena on, että jatkossa aineisto päivitetään tutkijoiden käyttöön noin puolivuosittain. Aineistosta ja sen avaamisen tavoitteista kertovat Kuluttajatutkimuskeskuksen tutkijat Lagus, Pantzar ja Ruckenstein näin[2]:

> ”Lähtökohtanamme on ollut, että Suomi24-aineiston avaaminen tukee yhteiskuntatieteellisen tutkimuksen uudistumista sekä uudenlaisen osaamisen ja yhteistyömuotojen rakentumista. Suomi24 on jaettu varanto, jonka ympärille tutkijat voivat kerääntyä. Keskusteluaineisto on ajallisesti pitkäkestoinen ja kansainvälisestikin poikkeuksellisen laaja sosiaalisen median aineisto. Viestejä on yli 70 miljoonaa ja niihin on tallentunut 15 vuoden ajalta suomalaista keskustelua.
>
> Suomi24 aineistoa luonnehtii asiakeskeisyys. Se perustuu vähemmän oman identiteetin rakentamiseen ja kaveriverkostoihin kuin esimerkiksi Facebook. Keskustelua vievät eteenpäin jaetut kiinnostuksenkohteet, esimerkiksi harrastukset, sukupuolinen suuntautuneisuus, perheen perustaminen tai terveyteen liittyvät ongelmat.”

1.1 SUOMI24 INFORMAATIOARKKITEHTUURI

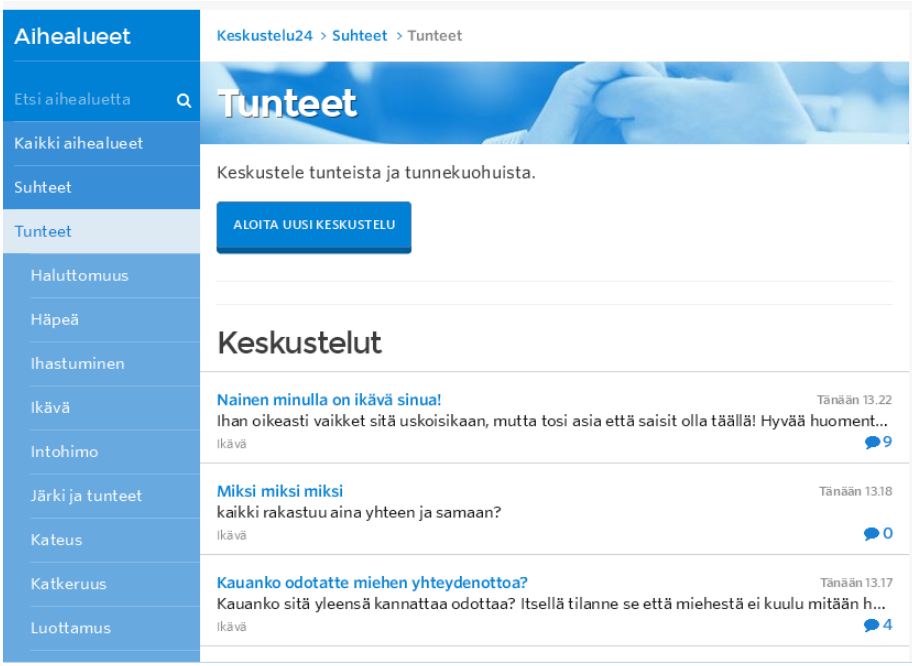
Aineiston syntyyn vaikuttavat paitsi käyttäjät ja heidän kiinnostuksensa ja odotuksensa, myös keskustelufoorumin informaatioarkkitehtuuri. informaatioarkkitehtuuri 1) ohjaa palvelun käyttäjien toimintaa ja 2) johtaa käyttäjien valikoitumi-

1 http://www.suomi24.fi

2 Krista Lagus, Mika Pantzar, Minna Ruckenstein (2015). Keskustelun tunneaallot – Suomi24-hanke, *Tieteessä Tapahtuu*, Vol. 33, Nro 6, s. 39–41.

seen. Käyttäjät, jotka eivät pidä esimerkiksi käyttöliittymän muutoksesta, hylkäävät foorumin. Toisaalta käyttöliittymän uudistuminen voi tehdä keskustelufoorumin hyväksyttäväksi ja käytettäväksi uusille kävijöille. Foorumin hierarkkinen alarakenne, sen muutokset, moderointikäytännöt sekä automaattiset suodattimet ohjaavat keskustelujen käynnistymistä, muovautumista ja päättymistä. Lisäksi keskustelujen syntyyn vaikuttaa ympäröivä digitaalinen kulttuuri laajemmin. Tähän sisältyvät kilpailevat sosiaalisen median välineet ja näissä tapahtuvat muutokset. Tässä raportissa emme kuitenkaan käsittele laajempaa digitaalisen kulttuurin kontekstia, vaan keskitymme Suomi24-foorumiin. Tietoja aineiston synnystä on saatu haastattelemalla Allerin työntekijöitä, esimerkiksi foorumin teknisiä toteuttajia ja moderaattoreita.

Suomi24-aineisto koostuu kommenteista, jotka joko käynnistävät keskusteluketjun tai kommentoivat jo olemassa olevaa keskusteluketjua. Kun palvelun kävijä haluaa aloittaa ketjun uudesta aiheesta, hän tulee samalla valinneeksi *keskustelualueen*, jonne sen sijoittaa. *Päätason* teema-alueet jakautuvat hierarkkisesti pienemmiksi, jopa viidenteen tai kuudenteen *alatasoon*. Näiden jälkeinen alin taso on *keskusteluketju*, jossa viestit ovat peräkkäin ja jossa vastaukset voivat kommentoida ketjun aiempia viestejä ja sisältää lainauksia näistä. Uusia keskustelualueita muodostaa Allerin henkilökunta. Palstoja luodaan tarpeen mukaan ja käyttäjiltä tulleiden ehdotusten pohjalta voidaan perustaa uusi palsta, jos sitä pidetään tarkoituksenmukaisena.



Kuva 1. Esimerkki keskusteluketjuista *Tunteet*-alueella

Kuvassa 1 on kuvankaappausesimerkki Suomi24-palvelun aihehierarkiasta. Vasemmalla sijaitsevasta palkista voidaan nähdä aihehierarkian rakenne ylös- ja alaspäin: Tunteiden yllä hierarkiassa on päätason alue *Suhteet,* ja sen alla puolestaan *Haluttomuus*, *Häpeä*, *Ihastuminen*, *Ikävä*, *Intohimo*, *Järki ja Tunteet*, *Kateus*, *Katkeruus* ja *Luottamus*. Keskusteluista näkyy ketjun otsikko, aloituskommentin alku ja kommenttien lukumäärä ketjussa sekä ketjun viimeisimmän kommentin ajankohta. Viimeksi kommentoitu ketju tai uusin aloitus on sivulla ylimpänä. Niin Suomi24-foorumin etusivulla kuin hierarkian alemmillakin tasoilla nostetaan esiin kunakin hetkenä suosituimmat keskusteluketjut ja keskustelualueet. Tämän vuoksi keskusteluihin osallistuminen painottuu usein edellisten päivien aktiivisiin ketjuihin.

Keskusteluketjun aloittaja antaa ketjulle otsikon, kun taas yksittäisillä kommenteilla ei ole otsikkoja. Toisaalta aineistossa on myös otsikoimattomia viestejä, koska aina otsikkokenttää ei ole käytetty. Yksittäisen ketjun kommenttimäärän maksimi on 500 kommenttia. Maksimimäärä on tekninen raja jonka tultua täyteen ketju sulkeutuu[3]. Samalla otsikolla voi tällöin halutessaan aloittaa uuden ketjun. Ketjuja voidaan sulkea myös moderaattorien päätöksellä, tai ennalta määritellyn epäaktiivisuusajan tullessa täyteen.

Kuvassa 2 on kuvankaappausesimerkki yhden viestin ketjusta keskustelualueella *Kaivinkoneet*. Keskusteluhierarkia näkyy viestin yläpalkissa seuraavasti: *Keskustelu24 > Ajoneuvot ja liikenne > Työkoneet ja raskas liikenne > Kaivinkoneet > äkeri ew200 hydrauliikka ja sähkökaaviot*. Viimeinen taso on ketjun otsikko, jonka ketjun aloittaja on nimennyt. Kuvassa näkyy myös mainos sekä kommenttikenttä, jolla ketjun keskusteluun voi osallistua, sekä *Ilmianna*-nappi, jonka avulla voi raportoida sisältöjä moderaattorille. Käyttöliittymän yksityiskohdat, kuten eri elementtien viemä tila ja niiden sijoittelu, voivat vaikuttaa suurestikin lukija- ja kirjoittajakokemukseen, ja sitä kautta aineiston syntyyn. Esimerkiksi Kuvan 2 mainoksen sijoittelu, ja sen viemä suuri tila ruudulla, voi vaikuttaa kävijän kykyyn hahmottaa keskustelun jatkumo ja ketju, johon hän on kommentoimassa. Ilmianna-napin sijoittelu ja sen saama huomio taas vaikuttaa siihen, kuinka herkästi ihmiset raportoivat häiritsevistä sisällöistä. Tällä taas on suora yhteys foorumin moderointiin ja sitä kautta keskustelujen koettuun laatuun.

Foorumin käyttöliittymä, mukaan lukien mainosten sijoittelu, on päivittynyt vuosien varrella; kirjoitushetkellä viimeisin päivitys on tehty 2015 tammikuussa. Allerilla on talletettuna kaikkien eri käyttöliittymäversioiden kuvat nettiarkistossa[4]. Tätä raporttia varten käyttöliittymäversioita ei kuitenkaan ole kartoitettu tai pohdittu niiden vaikutusta aineiston syntyyn.

3 Moderaattorien poistamat viestit eivät kartuta teknistä laskuria, joka vaikuttaa ketjun sulkemiseen 500 viestin tultua täyteen.

4 https://web.archive.org/web/*/suomi24.fi

Mobiililaitteilla tapahtuvan käytön helppous vaikuttaa laitteiden yleistyessä foorumien käyttöön ja käyttäjien osallistumiseen. Allerin mukaan Suomi24-foorumia ei ole erityisesti optimoitu mobiililaitteille, ja mobiilikäyttö on vähäistä. Käytettävyyttä on mahdollista tarkastella myös esteettömyyden näkökulmasta. Internet-vertaiskeskustelufoorumien teknisen toteutuksen käytettävyyttä, palvelun luonnetta ja käyttäjien kokemuksia on aiemmin kartoitettu RAY:n rahoittamien järjestöjen ylläpitämien vertaistukifoorumien osalta (Lagus et al, 2013)[5].

Kuva 2. Viesti keskustelualueella *Kaivinkoneet*

5 Lagus, Krista ; Saari, Juho ; Haaranen, Lassi ; Styrman, Tuula ; Tiainen, Ilkka. Kohti käyttäjien vertaisnettiä: Esikuva-analyysi ja uusia avauksia vertaistuen verkkopalveluista osana järjestöjen palvelutarjontaa. Raha-automaattiyhdistys, 2013.

1.2 SUOMI24 KÄYTTÄJÄT: KÄVIJÄT, KIRJOITTAJAT JA LUKIJAT

Foorumin *kävijöitä* ovat sekä sen *kirjoittajat* että *lukijat*. Kirjoittajat voivat olla rekisteröityneitä, jolloin kukaan muu ei voi käyttää heidän varaamaansa nimimerkkiä. Nimimerkin rekisteröiminen edellyttää toimivaa sähköpostiosoitetta. Palveluun voi kirjoittaa viestejä myös rekisteröitymättä, jolloin kirjoittaja voi viestin kirjoittamishetkellä valita jonkin nimimerkin. Yhdellä kävijällä voi näin ollen olla monta eri nimimerkkiä, ja toisaalta samalla rekisteröimättömällä nimimerkillä voi kirjoittaa useampi eri henkilö. Rekisteröityneitä kävijöitä koskevat käyttäjätiedot tai kävijöiden IP-osoite, josta viestit tulevat, eivät sisälly tutkimuskäyttöön tarjolla olevaan aineistoon. Aineisto ei näin ollen sisällä demografista yksilöintiä mahdollistavia tietoja.

Palvelun käyttöprofiileja on useanlaisia. Merkittävä osa kävijöistä saapuu palveluun Google-hakujen kautta. *Satunnaiset selaajat* saapuvat palveluun vain lukeakseen yhden tai useampia viestejä, ja jatkavat sitten matkaansa. Toista ääripäätä edustavat *aktiivikirjoittajat*, jotka viettävät päivittäin aikaa Suomi24:ssä, usein omilla suosikkialueillaan keskustellen. Osa aktiivikirjoittajista myös käyttää palvelua omana portaalinaan[6] verkkoon.

Aineistossa on mukana tieto siitä, montako kertaa tietty keskusteluketju on *näytetty* jollekulle kävijälle. Näyttökertoja ei lasketa viesteittäin vaan ketjua kohti. Tieto näyttökerroista on luonteeltaan päivittyvä, ja riippuu ajankohdasta jolloin aineisto on tallennettu Allerin tietokannasta. Sama aineisto myöhemmin kerättynä sisältää korkeampia näyttökertoja suurelle osalle ketjuja. Näyttökertojen kertymä on luonnollisesti myös suurempi vanhemmille ketjuille, koska ne ovat ehtineet saada esimerkiksi Google-hakujen osumia pidempään kuin uudemmat ketjut.

1.3 MODEROINTI

Kommentteja ja uusien ketjujen avauksia voidaan poistaa palvelusta moderaattorin toimesta. Moderointi tapahtuu jälkikäteen pääasiassa kävijöiltä tulleiden poistopyyntöjen perusteella, joihin Allerin palveluksessa olevat moderaattorit reagoivat pääsääntöisesti 1-2 vuorokauden sisällä. Keväällä 2016 Suomi24 keskusteluun tuli päivittäin viidestoista tuhannesta kahteenkymmeneen tuhanteen viestiä. Näistä viesteistä moderaattorit käsittelivät noin tuhat viestiä eli joko poistivat tai jättivät ne palstalle. Moderaattorin poistamasta kommentista jää aineistoon ja palveluun näkyviin tieto kommentin olemassaolosta ja ajankohdasta: kommentin teksti on

6 Portaali-termin kuvaus: https://fi.wikipedia.org/wiki/Portaali_%28internet%29 Suomi24 perustettiin alun pitäen internet-portaaliksi, ja vaikka tämä palvelumuoto ei olekaan enää Internetissä kovin yleinen, jotkut Suomi24-kävijät käyttävät palvelua edelleen tällä tavalla.

poistettu ja korvattu poistamisesta kertovalla virkkeellä. Moderaattorien poistamat tekstit eivät sisälly tutkimusaineistoon, sen sijaan metatiedot kuten kommentin ajankohta näkyvät foorumin käyttäjille ja ovat myös osa tutkimusaineistoa.

1.4 SUOMI24 YHTEISÖINÄ

Suomi24:n kirjoittajia voi ajatella yhtenä verkkoyhteisönä, jonka sisällä on lukuisia pienempiä yhteisöjä. Yhteisöjä muodostuu määriteltyjen keskustelupalstojen mukaan. Toisaalta keskustelijat voivat myös vaikuttaa siihen miten palstoja muodostetaan viemällä keskustelua tiettyyn suuntaan tai pyytämällä uusia palstoja ylläpidolta. Palstoilla voi ajatella olevan oma elinkaarensa. Ne kuolevat, kun ne eivät enää houkuta keskustelijoita.

Palstoilla keskustellaan yhteisistä kiinnostuksenkohteista tai vakaumuksista, saman elämäntilanteen jakavien tai samalla paikkakunnalla asuvien kanssa. Vilkkaita keskusteluja käydään esimerkiksi ikävästä ja lemmikkieläimistä: koiraihmiset juttelevat keskenään ja kissaihmiset keskenään. Osalle keskustelu on oman yrityksen tai poliittisen näkökulman edistämistä, toisille puhdasta ajanvietettä.

Suomi24:n keskustelijoiden anonyymius on näkökulmasta riippuen joko keskustelua rikastavaa tai rajoittavaa. Anonyymi keskustelu on luonteeltaan kevyttä, palstalla piipahdetaan jakamassa näkemyksiä ja tunnereaktioita. Toisaalta anonyyminä voi helpommin kertoa intiimeistä asioista. Anonyymius ei estä sitä, että ihmiset tulevat tutuiksi toisilleen. Anonyymi kirjoittaja voi käyttää yhtä tai useita nimimerkkejä. Oikeaa henkilöllisyyttä ei tiedetä, mutta kirjoittaja tunnistetaan ja hänen ympärilleen voi muodostua tiivis yhteisö. Tiiviitä yhteisöjä ovat esimerkiksi lesbojen ja 70-vuotiaiden palstat. Joskus yhteisö on niin tiivis, että uusia ja yhteisön ulkopuolisiksi koettuja keskustelijoita häädetään aktiivisesti pois palstalta.

Moderoinnin näkökulmasta palstat voivat näyttäytyä siisteinä tai helposti villiintyvinä. Siistit palstat, esimerkiksi matematiikkaa tai astrofysiikkaa käsittelevät palstat eivät kaipaa ylläpidon tukea. Viestejä tulee päivittäin, mutta keskustelut pysyvät asiallisina. Harrastuksiin liittyvä keskustelu, esimerkiksi ilmailupalsta, pysyy myös moderointinäkökulmasta pääosin puhtaana. Ajoittain keskustelu kuitenkin kiihtyy ja moderointi joutuu siihen puuttumaan. Pienten paikkakuntien palstat, työttömyyteen, potkuihin tai alkoholiin liittyvät keskustelut roihahtavat helposti asiattomiksi. Palstoilla on trollausta, niissä ärsytetään tahallaan ja provosoidaan muita keskustelijoita.

2. TUTKIJAN KÄYTTÖLIITTYMÄT AINEISTOON

Suomi24-aineisto on tallennettu Kielipankkiin[7]. Kielipankki on keskeinen kansallinen teksti- ja puheaineistojen palvelukokonaisuus, jota ylläpitää teknisesti CSC – Tieteen tietotekniikan keskus[8]. Kielipankkia operoi FIN-CLARIN infrastruktuuri, joka taas on osa kansainvälistä CLARIN ERIC infrastruktuuria. Kielipankki tarjoaa sekä aineistojen käyttömahdollisuuden että niiden käyttöön soveltuvia ohjelmistoja tehokkaassa laiteympäristössä. Joidenkin aineistojen käyttäminen edellyttää rekisteröitymisen tai luvan hakemisen. Kielipankki hoitaa käyttäjien tunnistamisen korkeakoulujen ja tutkimuslaitosten yhteistä Haka-käyttäjätunnistusjärjestelmää käyttäen.

Suomi24-aineistoa on tarjolla käyttäjille toistaiseksi kolmessa eri muodossa:

1. Tutkimustarkoituksiin, rekisteröityneille yliopistokäyttäjille aineisto on tarjolla suurena tiedostona, joka on ladattavissa Kielipankista joko VRT- tai JSON-formaatissa. Aineiston käyttö edellyttää sähköisesti allekirjoitettua lisenssiä, jossa sitoudutaan käyttämään aineistoa vain tutkimustarkoituksiin, ja olemaan luovuttamatta sitä eteenpäin. Ladattavaa aineistoa päivitetään Kielipankkiin[9]. Aineiston voi ladata osoitteesta https://korp.csc.fi/download/Suomi24/
2. Kaikille käyttäjille sanahakuina ns. *Korp*-käyttöliittymän kautta, joka toimii tavallisissa www-selaimissa. Käyttö ei vaadi rekisteröitymistä. Osoite https://korp.csc.fi/#?corpus=s24 sisältää Suomi24-aineistonäytteen. Koko Suomi24-aineisto koostuu useasta osakorpuksesta. Ne saa käyttöönsä Korpin valikkonäkymästä, jossa valitaan hakuun mukaan tulevat korpukset.
3. Ajantasaisin aineisto on Allerin API:n kautta haettavissa JSON-formaatissa. API:n käyttö edellyttää erillistä sopimusta suoraan Aller Oy:n kanssa.

2.1 RAAKADATAN JSON –TALLENNUSMUOTO

Suomi24-aineisto voidaan ladata yhtenä (tai useampana) suurena tiedostona käyttäjän koneelle JSON-muodossa. JSON on XML-tyyppinen sisäkkäisiä rakenteita sisältävä tallennusmuoto, jossa data kuvataan attribuutti-arvo-pareina.[10] JSON-

7 https://www.kielipankki.fi/

8 http://www.csc.fi/

9 Tavoitteena on, että aineisto päivittyisi Kielipankkiin noin puolen vuoden välein.

10 JSON-tallennusmuodon kuvaus Wikipediassa https://en.wikipedia.org/wiki/JSON

muodossa olevat tietueet sisältävät kaiken tutkimuskäyttöön annetun informaation (toisin kuin VRT-muoto, joka esitellään seuraavassa kappaleessa). Tallennusmuoto on sama, joka saadaan Allerin API:n[11] kautta. JSON-formaatti perustuu kenttiin ja kenttien erottimiin, jotka sisältävät muuttujia ja niiden arvoja. Kentän arvona voi olla myös lista, mahdollistaen sisäkkäiset rakenteet.

Eräs sattumanvaraisesti valittu ketju[12] on esitetty JSON-muodossa seuraavasti:

{”body”: ” Mukavasti on pukannut uusia S6 ja A8 Audeja liikenteeseen t\u00e4\u00e4ll\
u00e4 Turun seudulla!!! ”, ”closed_reason”: ”inactivity_time_reached”, ”views”: 308, ”deleted”: false, ”topics”: [{”title”: ”Ajoneuvot ja liikenne”, ”topic_id”: 2}, {”title”: ”Autot”, ”topic_id”: 6254}, {”title”: ”Automerkit”, ”topic_id”: 1109}, {”title”: ”Audi”, ”topic_id”: 3256}], ”title”: ”Audia pukkaa….”, ”comments”: [{”body”: ” pukkas 70:t\u00e4 luvulla Ladaa liikkuu jyv\u00e4skyl\u00e4ss\u00e4! ”, ”quote_id”: 0, ”deleted”: false, ”created_at”: 1207988326000, ”comment_id”: 29898190, ”anonnick”: ”krigg”, ”thread_id”: 5604224, ”parent_comment_id”: 0}], ”anonnick”: ”SS66AA88”, ”thread_id”: 5604224, ”closed”: true, ”created_at”: 1207910044000}

Esimerkiksi yllä ”anonnick”: ”krigg” muuttuja-arvo-pari ilmaisee, että testin kirjoittanut anonyymi nimimerkki on ”krigg”. Anonyymi nimimerkki lienee kirjoittaja, joka ei ole rekisteröimällä suojannut nimimerkkiään vaan on vain kirjoittanut sen nimimerkkikenttään kommenttia lähettäessään. Ääkkösten ja muiden erikoismerkkien koodaus (esim. ä = \u00e4) noudattaa C/C++/Java/Python-ohjelmointikielissä käytetty Unicode-merkintätapaa.

JSON-formaatissa oleva *tietue* sisältää yhden keskusteluketjun tiedot. Kun se rivitetään huomioiden erottimia, kuten kaarisulut, hakasulut ja pilkut, ja sisennetään sisempien sulkujen kohdalla, saadaan tulokseksi helpommin luettava muoto, josta tietueen loogisen rakenteen hahmottaminen on huomattavasti helpompaa:

{”body”: ” Mukavasti on pukannut uusia S6 ja A8 Audeja liikenteeseen t\u00e4\u00e4ll\
u00e4 Turun seudulla!!! ”,
“closed_reason”: “inactivity_time_reached”,
“views”: 308,
“deleted”: false,
“topics”: [{“title”: “Ajoneuvot ja liikenne”, “topic_id”: 2},
{“title”: “Autot”, “topic_id”: 6254},
{“title”: “Automerkit”, “topic_id”: 1109},
{“title”: “Audi”, “topic_id”: 3256}],

11 Allerin API:n dokumentaatio osoitteessa https://www.suomi24.fi/static/ops/API.html

12 http://keskustelu.suomi24.fi/t/5604224/audia-pukkaa----

“title”: “Audia pukkaa….”,
“comments”:
[{“body”: “ pukkas 70:t\u00e4 luvulla Ladaa liikkuu jyv\u00e4skyl\u00e4ss\u00e4! ”,
“quote_id”: 0, “deleted”: false, “created_at”: 1207988326000, “comment_id”: 29898190,
“anonnick”: “krigg”, “thread_id”: 5604224, “parent_comment_id”: 0}],
“anonnick”: “SS66AA88”,
“thread_id”: 5604224,
“closed”: true,
”created_at”: 1207910044000}

Yllä siis näyttäisi olevan anonyymin nimimerkin *SS66AA88* käynnistämä ketju, johon on tullut yksi kommentti nimimerkiltä *krigg*. Ketjun aloitusviesti on ensimmäisen *body*-kentän arvona: *Mukavasti on pukannut uusia S6 ja A8 Audeja liikenteeseen täällä* Turun seudulla*!!!*. Ketju on näytetty 308 kertaa, ja sitä ei ole poistettu moderaattorin toimesta. Aihehierarkia ilmaistaan *topics*-kentässä. Esimerkin keskusteluketju on aihehierarkian neljännellä tasolla, *Audi*-alaryhmässä. Ketjun aloitusviestin otsikko on ”*Audia pukkaa*”. Aloituksen saamat kommentit aikaleimoineen ovat *comments*-kentän arvona, sisempänä listarakenteena. Ketju on suljettu, ja sulkemisen syyksi on kerrottu että ketju on ollut epäaktiivinen määrätyn ajan. Luontihetki on UNIX-aikaleima[13] millisekunteina: ”1207910044000” kertoo että ketju eli sen käynnistänyt kommentti on luotu 11.4.2008 klo 13:34. Ketjun tunniste ”thread id” on muunnettavissa linkiksi Suomi24-foorumin ketjuun seuraavasti: http://keskustelu.suomi24.fi/t/thread_id .

2.2 KIELELLISESTI ANALYSOITU VRT-TALLENNUSMUOTO

JSON-formaatin lisäksi koko aineisto voidaan ladata käyttäjän koneelle myös niin kutsutussa VRT-muodossa. Alkuperäisen Suomi24-foorumin aineistoa on tässä täydennetty sanojen ja virkkeiden kielellisillä, niin sanotuilla morfosyntaktisilla analyyseillä. Morfosyntaktiset analyysit aineistoon on tuottanut Kielipankki, FIN-CLARIN.

Verticalized text eli VRT on rivipohjainen tapa esittää tekstejä koskevia analyysitietoja. Siinä tekstien lauseet on jaettu yksi sana (*token*) riville. Sanojen lisäksi myös välimerkit ovat omilla riveillään. Sanan jälkeen samalla rivillä on sarkaimella erotettuna joukko sanaan liittyviä attribuutteja, kuten morfosyntaktisen analyysin

13 UNIX-aikaleiman muuntamiseen ihmisen luettavissa olevaksi ajankohdaksi löytyy valmiit työkalut useista ohjelmakirjastoista joilla tämän tyyppisiä kokonaisaineistoja analysoidaan.

automaattisesti[14] tuottamia luokituksia. Näitä ovat mm. sanan perusmuoto, sanaluokka, mahdolliset sijamuodot, jne. Suomi24-aineiston VRT-muoto on laadittu Korp-työkalua varten ja noudattaa sen edellyttämää formaattia[15].

VRT-muodossa aineisto on kirjoitushetkellä tarjolla kahdessa eri koossa: 1,5 gigan versio, jossa on murto-osa koko Suomi24-aineiston dokumenteista keräyshetkeen (toukokuu 2015) mennessä, ja 28 gigan versio, jossa on kaikki Allerilta saatu materiaali. Aineistot on kompressoitu zip-muotoon.

Aineiston määristä on kerrottu seuraavasti[16]:

texts:	48,423,611	(counting <text)
sentences:	231,373,748	(counting <sentence)
tokens:	2,385,073,226	(counting ^[^<] [corrected 2015-11-02])

Texts viittaa dokumenttien eli kommenttien määrään, *sentences* lauseisiin ja *tokens* sanoihin tai välimerkkeihin. Aineiston esikäsittelystä kerrotaan aineiston yhteydessä seuraavaa:

> The tokenized version was created by Aleksi Sahala. Annotation process was then carried out by Jussi Piitulainen (using CSC's Taito cluster). The morpho-syntactic analysis was produced with the Turku Dependency Parser.

Alla on satunnaisin perustein valittu esimerkki VRT-tallennusmuodossa olevasta kommentista ”Onko totta, että kateus vie kalat vedestä?”. Kommentti on erään ketjun[17] loppupuolelta.

```
<text title=”Se parhaiten nauraa” sect=”Ryhmät” sub=”60 plus” user=”O - O”
date=”14.10.2013” time=”19:59” datefrom=”20131014” dateto=”20131014” nid=”11783377”
cid=”64000085” urlboard=”http://keskustelu.suomi24.fi/t/11783377” urlmsg=”http://keskus-
telu.suomi24.fi/t/11783377#comment-64000085”>
<paragraph id=”2091380”>
<sentence id=”9935405”>
Onko    1    olla    V    PRS_Sg3|VOICE_Act|TENSE_Prs|MOOD_Ind|CLIT_
Qst|CASECHANGE_Up    2    cop    _
totta    2    tosi    A    NUM_Sg|CASE_Par|CMP_Pos    0
ROOT    _
,    3    ,    Punct    _    6    punct    _
että    4    että    C    SUBCAT_CS    6    complm    _
kateus    5    kateus    N    NUM_Sg|CASE_Nom    6    nsubj    _
```

14 Kielellinen analyysi on tehty automaattisilla työkaluilla, ja sisältää vääjäämättä jonkin verran virheitä. Erityisen haasteen analyysille tuottavat puhekieliset ilmaukset sekä kirjoitusvirheet.

15 https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiOhjeetKorpFormaatti

16 https://korp.csc.fi/download/Suomi24/Suomi24-2015-10-29_VRT_README.txt

17 http://keskustelu.suomi24.fi/t/11783377/se-parhaiten-nauraa

```
vie 6 viedä V PRS_Sg3|VOICE_Act|TENSE_Prs|MOOD_Ind 2 csubj-cop _
kalat 7 kala N NUM_Pl|CASE_Nom 6 dobj _
vedestä 8 vesi N NUM_Sg|CASE_Ela 6 nommod _
? 9 ? Punct _ 2 punct _
</sentence>
</paragraph>
</text>
```

Esimerkistä näkyy, kuinka jokainen sana ja välimerkki on VRT-muodossa tallennettuna omalla rivillään. Virkkeen sanat on lisäksi merkattu mm. sanan perusmuodolla, sanaluokalla ja erilaisilla määreillä. Määreet kertovat sanan morfologisista ja lauseenjäsennykseen liittyvistä ominaisuuksista. VRT-tallennusmuoto sisältää siis paitsi alkuperäisen kommentin tekstin, myös morfosyntaktisen analyysin tekstistä. Toisaalta VRT-tallennusmuoto ei toistaiseksi sisällä osaa JSON-formaatissa olevista tiedoista: esimerkiksi kommentin kategoriahierarkiasta VRT-esitysmuotoon on tallennettu vain kaksi ylintä hierarkiatasoa, ja kirjoittajan nimimerkkiä ei tässä muodossa ole tallennettu[18].

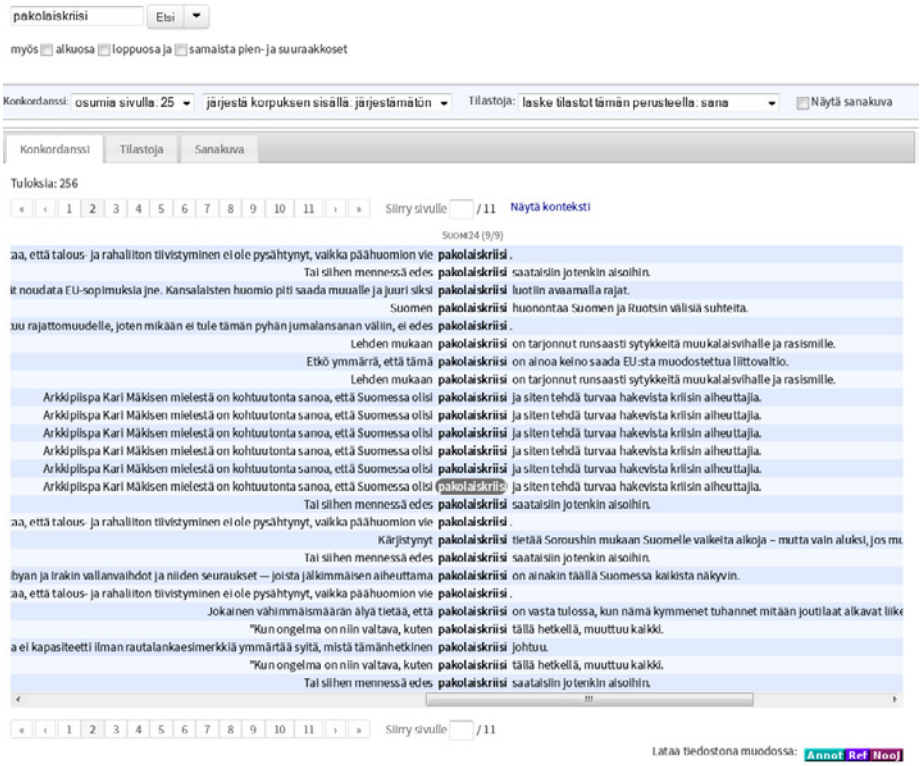
2.3 KORP-HAKUKÄYTTÖLIITTYMÄ

Suomi24-aineistoon voi kuka tahansa tehdä hakuja Korp-hakukäyttöliittymän kautta ilman rekisteröitymistä. Korp on interaktiivinen, www-käyttöliittymän kautta toimiva kieliopillisesti jäsennettyjen tekstiaineistojen hakutyökalu[19]. Korp-käyttöliittymä käyttää pohjanaan VRT-tallennusmuodon sisältämiä tietoja, eli se sisältää tiedot kielen sanaluokista ja lauseenjäsennyksistä.

Korpissa voi tehdä valitsemaansa tekstiaineistoon (tai aineistoihin) kohdistuvan haun sanalla, sen osalla, tai vaikkapa sanan perusmuodolla tai sanaluokalla. Hakua voi rajata dokumentin metatietojen perusteella. Myös usean peräkkäisen sanan hakuja, tai vaihtoehtoisten sanojen muodostaman joukon hakuja on mahdollista muodostaa. Osumat näytetään ensisijaisesti ns. *konkordanssina* eli lyhyinä lausekonteksteina, joissa hakutulokset ovat ruudulla allekkain, haettu sana keskitettynä (Kuva 3). Konkordanssinäyttö on perinteinen kielitieteen piirissä käytetty tapa tarkastella sanojen lyhyitä lausekonteksteja.

18 FIN-CLARINilta saadun tiedon mukaan myöhemmin päivittyvät aineistoversiot tulevat sisältämään loputkin aineistossa mukana olevat kentät myös aineiston VRT-formaatissa.

19 Korp-käyttöohje ja tietoa työkalusta löytyy osoitteesta https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiOhjeetKorp



aa, että talous- ja rahaliiton tiivistyminen ei ole pysähtynyt, vaikka päähuomion vie	**pakolaiskriisi**	.
Tai siihen mennessä edes	**pakolaiskriisi**	saataisiin jotenkin aisoihin.
it noudata EU-sopimuksia jne. Kansalaisten huomio piti saada muualle ja juuri siksi	**pakolaiskriisi**	luotiin avaamalla rajat.
Suomen	**pakolaiskriisi**	huonontaa Suomen ja Ruotsin välisiä suhteita.
uu rajattomuudelle, joten mikään ei tule tämän pyhän jumalansanan väliin, ei edes	**pakolaiskriisi**	.
Lehden mukaan	**pakolaiskriisi**	on tarjonnut runsaasti sytykkeitä muukalaisvihalle ja rasismille.
Etkö ymmärrä, että tämä	**pakolaiskriisi**	on ainoa keino saada EU:sta muodostettua liittovaltio.
Lehden mukaan	**pakolaiskriisi**	on tarjonnut runsaasti sytykkeitä muukalaisvihalle ja rasismille.
Arkkipiispa Kari Mäkisen mielestä on kohtuutonta sanoa, että Suomessa olisi	**pakolaiskriisi**	ja siten tehdä turvaa hakevista kriisin aiheuttajia.
Arkkipiispa Kari Mäkisen mielestä on kohtuutonta sanoa, että Suomessa olisi	**pakolaiskriisi**	ja siten tehdä turvaa hakevista kriisin aiheuttajia.
Arkkipiispa Kari Mäkisen mielestä on kohtuutonta sanoa, että Suomessa olisi	**pakolaiskriisi**	ja siten tehdä turvaa hakevista kriisin aiheuttajia.
Arkkipiispa Kari Mäkisen mielestä on kohtuutonta sanoa, että Suomessa olisi	**pakolaiskriisi**	ja siten tehdä turvaa hakevista kriisin aiheuttajia.
Arkkipiispa Kari Mäkisen mielestä on kohtuutonta sanoa, että Suomessa olisi	**pakolaiskriisi**	ja siten tehdä turvaa hakevista kriisin aiheuttajia.
Arkkipiispa Kari Mäkisen mielestä on kohtuutonta sanoa, että Suomessa olisi	**pakolaiskriisi**	ja siten tehdä turvaa hakevista kriisin aiheuttajia.
Tai siihen mennessä edes	**pakolaiskriisi**	saataisiin jotenkin aisoihin.
aa, että talous- ja rahaliiton tiivistyminen ei ole pysähtynyt, vaikka päähuomion vie	**pakolaiskriisi**	.
Kärjistynyt	**pakolaiskriisi**	tietää Soroushin mukaan Suomelle vaikeita aikoja – mutta vain aluksi, jos mu
Tai siihen mennessä edes	**pakolaiskriisi**	saataisiin jotenkin aisoihin.
byan ja Irakin vallanvaihdot ja niiden seuraukset — joista jälkimmäisen aiheuttama	**pakolaiskriisi**	on ainakin täällä Suomessa kaikista näkyvin.
aa, että talous- ja rahaliiton tiivistyminen ei ole pysähtynyt, vaikka päähuomion vie	**pakolaiskriisi**	.
Jokainen vähimmäismäärän älyä tietää, että	**pakolaiskriisi**	on vasta tulossa, kun nämä kymmenet tuhannet mitään joutilaat alkavat liike
"Kun ongelma on niin valtava, kuten	**pakolaiskriisi**	tällä hetkellä, muuttuu kaikki.
a ei kapasiteetti ilman rautalankaesimerkkiä ymmärtää syitä, mistä tämänhetkinen	**pakolaiskriisi**	johtuu.
"Kun ongelma on niin valtava, kuten	**pakolaiskriisi**	tällä hetkellä, muuttuu kaikki.
Tai siihen mennessä edes	**pakolaiskriisi**	saataisiin jotenkin aisoihin.

Kuva 3. Korp-tuloksia hakusanalle ”pakolaiskriisi” konkordanssimuodossa

Esimerkissä keskellä näkyy viisi saman sisältöistä kontekstia. Tarkemmin perehdyttäessä huomataan, että ilmentymät ovat kaikki peräisin saman nimimerkin kommenteista, jotka on lisätty peräkkäisinä ajanhetkinä eri keskustelualueille. Konkordanssinäytön lisäksi Korpissa on mahdollista pyytää näkyville myös *kappalekontekstit*. Yhden näytön sisältämät tulokset on mahdollista tallentaa ruutu kerrallaan ja ladata omalle koneelleen. Hakujen tuloksia voi tarkastella ajallisesti, kirjoitushetkellä vuoden tarkkuudella, kohdasta Tilastoja / Näytä Trendidiagrammi.

Edistyneemmälle käyttäjälle on tarjolla tapoja joilla voi määritellä haun kohteeksi esim. sanapareja tai lauserakenteita. Haut voi myös tallettaa ja vertailla eri hakujen tuloksia keskenään.

Suomi24-aineiston lisäksi Korpin kautta voi hakea tekstejä muistakin aineistoista, mukaan lukien sanomalehti- ja kirja-aineistoja. Internet-aineistoista mukana on esimerkiksi nuorten suosima keskustelufoorumi Ylilauta. https://korp.csc.fi vie näkymään josta on mahdollista valita haun kohteena kulloinkin olevat aineistot.

2.4 AINEISTON VERSIOT JA PÄIVITTYMINEN

Suomi24-palveluun syntyy jatkuvasti lisää palvelun kävijöiden kirjoittamia tekstejä. Varsinaisiksi tutkimusaineistoiksi palvelusta haetaan ja valmistellaan täydennettyjä dataversioita resurssien niin salliessa. *Aineiston tutkimuskäyttöön siirtäminen* kattaa mm. aineiston keruun Allerin tarjoaman API:n kautta, kahdentuneiden viestien poistoa, aineiston virheiden korjaamista ja esikäsittelyä eri tavoin sekä sen täydentämistä kielellisillä analyyseillä.

Korp-aineiston keräsivät ja analysoivat FIN-CLARIN:in tutkijat. Ensimmäisen JSON-muotoisen version keräsi ja paketoi saataville tohtorikoulutettava Matti Nelimarkka. Myöhemmät VRT- ja Korp-versiot ovat syntyneet FIN-CLARIN:in työnä. Aineiston päivitysprosessia ja siihen liittyviä käytäntöjä ja ohjelmia kehitetään Citizen Mindscapes – Kansakunnan mielenliikkeet -konsortiohankkeessa. Työtä koordinoi Helsingin yliopiston Menetelmäkeskus. Toistaiseksi aineistoa on tarjolla tutkimuskäyttöön seuraavasti:

Ladattavan aineiston päivitykset:

- *Toukokuussa 2015* koko siihenastinen **Suomi24** aineisto JSON-formaatissa (tunniste Suomi24-2015-05-25_JSON), Aineisto on ajalta 2001 – toukokuu 2015. Aineisto tarjottiin Kielipankin kautta tutkijoille sekä muistitikulla Suomi24-hacakthonin osallistujille 29. – 30.5.2015.
- *Toukokuussa 2015* **Suomi24 2001 – 2014 (näyte)** -aineisto VRT-formaatissa (tunniste Suomi24-2015-04-02_VRT) tarjottiin Kielipankin kautta tutkijoille sekä muistitikulla Suomi24-hacakthonin osallistujille 29. – 30.5.2015.
- *Marraskuussa 2015* Kielipankkiin tarjolle tuli VRT-muodossa oleva **Suomi24** aineisto ajalta 2001 – kesäkuu 2015 (tunniste Suomi24-2015-10-29_VRT).

Korp-aineiston päivitykset:

- *Huhtikuussa 2015* aineiston osajoukko tuli tarjolle Korp-hakukäyttöliittymän kautta kaikille kiinnostuneille. Osajoukon laajuus oli tällöin arviolta 5-7 % siihenastisesta foorumin aineistosta, sisältäen noin 120 miljoonaa sanaa (tokens). Aineiston avaaminen julkistettiin 9.4.2015 ilmestyneellä tiedotteella. Aineisto löytyy valikossa ”Internet-keskusteluaineistoja” nimellä ”**Suomi24 2001–2014 (näyte)**”.
- *Joulukuussa 2015* Korpissa tuli tarjolle koko Suomi24-aineisto ajalta 1.1.2001–18.11.2015. Aineisto löytyy kohdasta ”Internet-keskusteluaineistoja” nimellä ”**Suomi24**”. Se on laajuutensa vuoksi jaettu yhdeksään osakorpukseen. Jako johtuu puhtaasti teknisistä syistä – koko Suomi24-korpusta käytettäessä tulisi käyttöliittymästä valita mukaan kaikki sen yhdeksän osiota.

VRT-muodon päivitys Korpiin kestää tavallisesti kauemmin kuin ladattavien JSON-tiedostojen päivitys, koska aineiston VRT-muotoon saattamisen jälkeen siitä lasketaan vielä Korpia varten erilaisia hakupalvelimen edellyttämiä indeksejä ja muita tietoja.

Aineistojen ja versioiden nimeäminen, päivityskäytännöt, sekä näitä toteuttavat prosessit ja ohjelmat hakevat vielä muotoaan. Tässä raportissa on omaksuttu Korp-käyttöliittymän nimeämiskäytäntö, jossa tämänhetkiseen kokonaisaineistoon viitataan nimellä **Suomi24** ja aineiston pieneen osajoukkoon joka aluksi oli tarjolla Korpin kautta nimellä **Suomi24 2001–2014 (näyte)**.

3. AINEISTON OMINAISUUKSIA

Tässä kuvaamme aineistoa erilaisista määrällisiä ja mitattavia näkökulmia käyttäen. Mittaluvut ja jakaumakuvat on tuotettu systemaattisella tavalla ohjelmallisesti lukujen osalta ja käyttäen Exceliä kuvaajien visualisoinnissa. Aineiston käsittelyssä tehdyt valinnat muun muassa esikäsittelyn osalta on pyritty kuvaamaan osana raporttia. [20]

3.1 AINEISTON KOKONAISMÄÄRÄT JA VUOSITTAISET MÄÄRÄT

Aineistojen määrät eri tallennusmuodoissa vaihtelevat. Yksi tapa tarkastella aineiston määriä on kommenttien kasvava id-indeksi, joka on suurimmassa toistaiseksi olevassa tallenteessa lähes 80 miljoonaa[21]. Toistaiseksi tarjolla oleva aineisto näyttäisi sisältävän kuitenkin enimmilläänkin alle 60 miljoonaa kommenttia. Korp-käyttöliittymän kautta kulloinkin tarjolla olevan aineiston määrä sanoina laskettuna on kerrottu hakukäyttöliittymässä itsessään: dokumenttimääriä ei Korp-käyttöliittymässä mainita. Korpissa on kuitenkin aina tarjolla jokin VRT-formaatissa esillä oleva aineisto.

Alla olevat laskelmat[22] aineiston määristä ja jakaumista on laskettu Kielipankissa JSON-tallennusmuodossa tarjolla olevasta aineistosta (tunniste Suomi24-2015-05-25_JSON).

Kielipankissa tarjolla olevasta aineistosta puuttuu materiaalia indeksimääriin verrattuna ainakin kolmesta syystä:

1. *Seksi-keskustelualueen kommentit*. Seksi-alue on tarkoitettu yli 18 vuotta täyttäneille käyttäjille, eivätkä sen kommentit olleet samanarvoisina tar-

20 Tiukat toistettavuuden kriteerit eivät aineistoraportin kuvaajien ja tulosten osalta toteudu. Toistettavuus edellyttäisi kaikkien aineiston käsittelyssä tarvittujen ohjelmapalasten ja ohjelmaskriptien julkaisemista. Tässä yhteydessä täydelliseen toistettavuuteen ei ole edes pyritty, vaan meille on riittänyt aineistokokonaisuuden hahmottelu. Suomi24-aineisto, kuten todellisuus, jota se heijastelee ja ikään kuin mittaa, on luonteeltaan alati muuntuva ja päivittyvä. Jopa aineiston lataushetki vaikuttaa aineiston sisältöön ja näin ollen jokaisen ketjun latausajankohta API:sta tulisi raportoida: myöhemmin saatavilla ei ole enää täsmälleen sama aineisto. Mittaluokiltaan luvut ja jakaumat pitänevät paikkansa riittävällä tarkkuudella senkaltaisten johtopäätösten tekemiseen, joita tällaisesta aineistosta on Suomi24-foorumin keskusteluja koskien mielekästä vetää.

21 Suurimmassa tallenteessa 21.5. saakka kerätyn JSON-aineiston viimeinen kommentti-indeksi on 79 162 786. Kommentti-indeksi ei pidä sisällään keskusteluketjujen avauksia vaan niillä on oma indeksinsä, jonka suurin arvo JSON-aineistossa on 13 595 215.

22 Laskelmat ja niiden tarvitsemat ohjelmointiskriptit on tehnyt Pasi Karhu (StudioMind Oy).

jolla Allerin API-liittymän kautta, joten ne eivät kuulu tutkimusaineistoon. Jatkossa aineisto on kuitenkin tarkoitus täydentää myös Seksi-keskustelualueen osalta.

2. *Moderaattorien poistamat kommentit*. Moderoitaviksi katsotut kommentit on poistettu kokonaan foorumilta – niistä on jäljellä vain otsikko- eli ns. header-tiedot (kuten kommentin ajankohta), sekä poistamisesta kertova kuuden sanan automaattinen vakiofraasi.
3. *Muista syistä puuttuvat kommentit*. Ajallisessa tarkastelussa on vaikuttanut siltä, että aineistosta puuttuu jonkin verran kommentteja joiltakin ajanjaksoilta. Täsmällinen syy ei ole selvillä, mutta mahdollisia syitä ovat esimerkiksi palvelun alhaalla olo ohjelmistopäivityksen takia, tietokantamuunnoksissa hävinnyt data tai virheet JSON-aineiston poimintaskripteissä. Eräs esimerkki puuttuvasta aineistosta ovat päivät 19.–20.5.2002 ja 6.12.2007 joiden kommentteja ei ole datasetissä lainkaan. Näinäkin päivinä on kuitenkin käytetty kommentti-ID-numeroita, joten foorumi on ollut toiminnassa silloinkin, vaikka kommentit eivät ole ilmeisesti tallentuneet tietokantaan. Aineistossa on havaittavissa jonkin verran myös aikaleimojen sekaantumista[23], mikä on nähtävissä myös Suomi24-foorumilla. Aikaleimojen sekaantuminen on lähes aina tietokannan teknisten tallennusvirheiden seuraus.

Taulukkoon 1 on koottu kokonaismäärät aineistosta, joka on kerätty API:n kautta 14.–21.5.2015 ja on tarjolla Kielipankissa tunnisteella ”Suomi24-2015-05-25_JSON”. Se sisältää 555 kpl erillisiä tiedostoja, joiden koot vaihtelevat välillä 5,8 MB ja 1035 MB. Tiedostojen yhteiskoko on 33,6 GB. Aineisto sisältää ketjut, jotka on tuotettu Suomi24:ään välillä 1.1.2001–15.5.2015. Aineistoon on ennen laskelmia tehty seuraavat poistot:

- Noin 500,000 viestiä joissa on sama kommentti-id. Nämä lienevät mukana aineistossa johtuen Allerin API:n kautta tehtävään lataukseen liittyvistä haasteista, eikä niitä voi pitää osana varsinaista aineiston korpusta.
- Noin 100,000 viestiä joiden kommentti-id on eri kuin toisessa kommentissa, mutta muut tiedot samoja, ja kommenttien aikaleimoissa eli lähetyshetkissä on vain lyhyt ero. Nämä katsotaan kommenttiduplikaateiksi, joita voi tallentua esimerkiksi käyttäjän tarkoituksettoman ”Lähetä”-painikkeen kaksoisklikkauksen takia. Tällainen viesti näkyy keskustelufoorumillakin kahteen kertaan, ja niitä voidaan pitää osana korpusta, koska ne ovat muiden käyttäjien nähtävissä foorumilla. Tämän raportin laskelmista kyseiset viestit on kuitenkin poistettu, koska käyttäjien toiminnan ja pyrkimysten

23 Aikaleimojen sekaantuessa keskustelun kommentit menevät palstan ketjussa keskenään epäjärjestykseen. Vaihtoehtoisesti keskustelun kommentit ovat ketjua palstalla lukiessa järjestyksessä, mutta aikaleiman perusteella näyttää siltä, että vastaus olisi kirjoitettu ennen kysymyksen esittämistä.

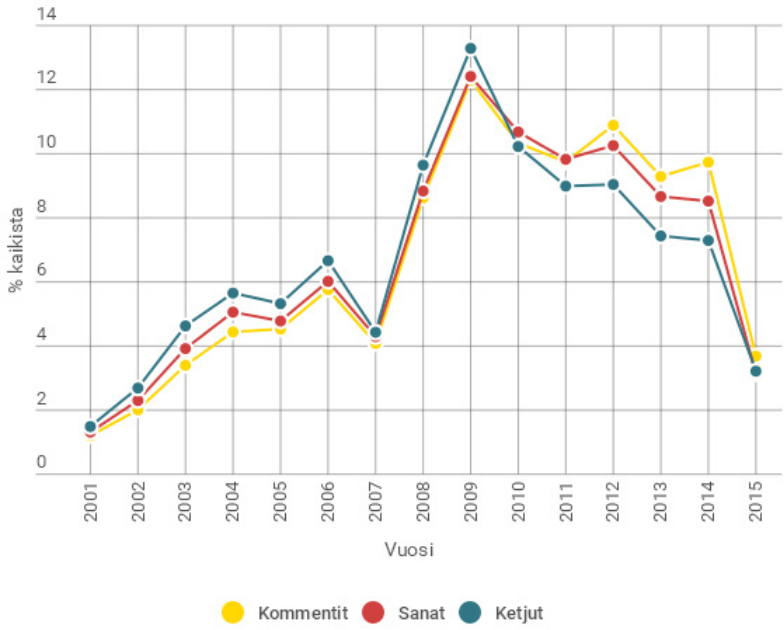
tutkimisen kannalta ne näyttäytyvät satunnaisena virhelähteenä, kohinana jonka pyrimme poistamaan.

Laskennassa ovat toisaalta mukana viestit jotka moderaattorit ovat poistaneet – nämäkin kun ovat käyttäjien lähettämiä viestejä.

Taulukko 1. Aineiston kokonaismäärät

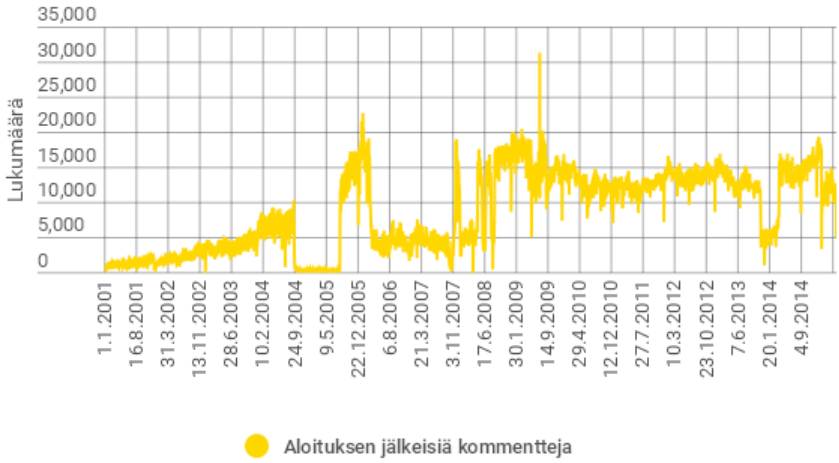
Sanoja yhteensä	2 188 829 886
Kommentteja yhteensä	53 401 214
Keskusteluketjuja yhteensä	6 805 028
Ketjujen näyttökertoja yhteensä	3 493 463 004

Näyttökertojen ja kommenttien välinen suhde on keskimäärin noin 64 näyttökertaa ketjulle jokaista siihen kirjoitettua kommenttia kohden (Taulukko 6). Tämä suhde kuitenkin vaihtelee suuresti aihepiireittäin. Lisäksi suhde on tyypillisesti keskimäärin suurempi vanhemmalla aineistolla, joka on ehtinyt kerätä näyttökertoja pidemmän ajan kuin tuoreemmat viestit.



Kuva 4. Kommenttien, sanojen ja ketjujen vuosittaiset prosenttiosuudet

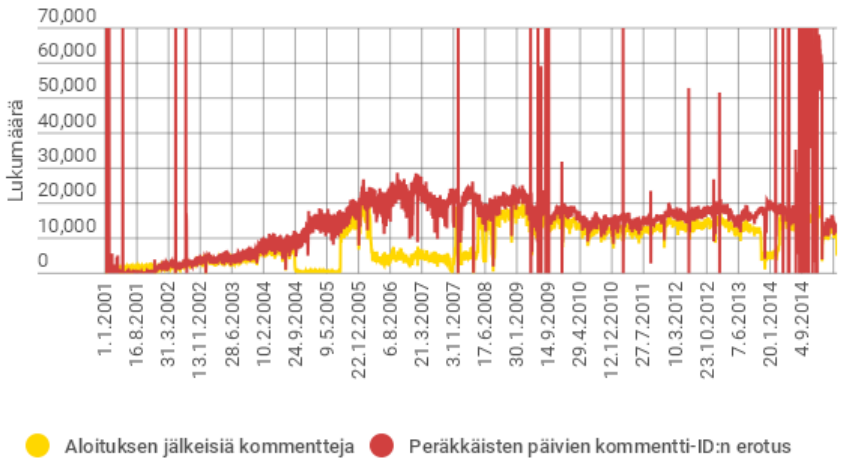
Aineiston määrällinen kehitys näyttäisi Kuvan 4 perusteella olevan vuodesta 2010 lähtien melko tasainen kommenttien määrissä tarkasteltuna. Yllättäviä poikkeamia ovat vuoden 2007 vaje ja huippuvuosi 2009. Tarkasteltaessa aineiston päivittäisjakaumaa sen selvittämiseksi mistä vajeessa voisi olla kyse, aineiston kommenttien kertymistahti osoittautui seuraavanlaiseksi (Kuva 5):



Kuva 5. Aineiston kertymistahti Suomi24-palstan historian aikana

Kuvan 5 aineiston kertymistahdin päivittäisjakaumaa tarkastelemalla siitä on nähtävissä kolme aineistovajetta. Ensimmäinen, pienehkö pidemmän aikavälin vaje on syntynyt loppuvuodesta 2004 loppuvuoteen 2005. Toisaalta vuoden 2007 pidemmän aineistovajeen kesto tarkentuu: se näyttäisi alkaneen vuoden 2006 alkupuolella ja jatkuneen kesäkuulle 2008. Kolmas, pienempi vaje on nähtävissä vuoden 2013 loppupuolella.

Kuten edellä todettiin, aineisto ei sisällä kaikkia Suomi24-palstalle kirjoitettuja kommentteja; siitä puuttuvat teknisistä syistä kadonneet kommentit, Seksi-alueen keskustelut sekä moderaattorien poistamat viestit. Koska kommenttien määrän kasvaessa myös kommentti-id:den määrä kasvaa, id:n kasvua tarkastelemalla voidaan arvioida yleisellä tasolla puuttuvien Seksi-alueen aineiston ja moderoitujen kommenttien kertymistahtia sekä teknisten ongelmien vaikutusta. Aineistosta tutkittiin siksi vielä kommentti-id:den kertymistahdin ja käytettävissä olevan aineiston kertymistahdin välinen ero.



Kuva 6. Kommenttien kertymistahti Suomi24-aineistossa a

Kuvassa 6 esitetyn id:iden karttumisen perusteella vaikuttaa siltä, että keskustelupalstalle olisi kaikkien kolmen esitetyn aineistovajeen aikana tullut huomattavasti enemmän kommentteja kuin mitä käytettävissä oleva aineisto sisältää. Tarkkaa syytä näihin vajeisiin ei kirjoitushetkellä selvittelyjenkään jälkeen tiedetä, mutta yleisesti yhtäkkiset muutokset aineistomäärissä seuraavat joko teknisistä ongelmista, koko sivustoa koskevista uudistuksista, joiden aikana palvelu ei välttämättä ole lainkaan käytettävissä, tai muutoksista joissain suosituissa aihealueissa tai palstoissa.

Tietoisuus aineiston ajallisista vajeista ja virheistä on tärkeää etenkin silloin jos halutaan tutkia juuri tietylle ajanjaksolle sijoittuvia yhteiskunnallisia tapahtumia ja niitä koskevaa vuoropuhelua. Edellä esitetyn analyysin perusteella Suomi24-aineisto ei välttämättä tarjoa apua esimerkiksi vuoden 2005 alkupuolen, tietyiltä osin lähes koko vuoden 2007, sekä loppuvuoden 2013 tapahtumien tutkimukselle.

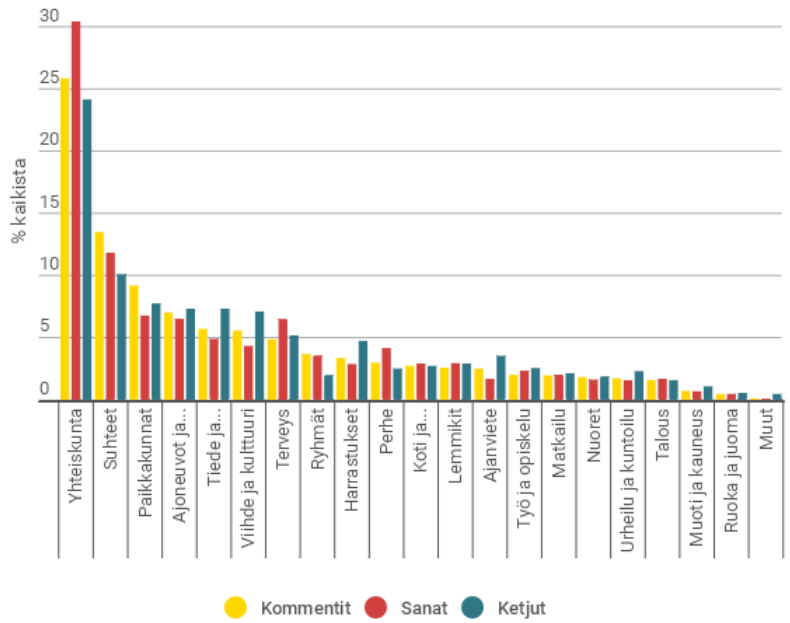
3.2 AINEISTON MÄÄRÄT ERI KESKUSTELUALUEILLA

Aihealueiden hierarkia tarjoaa luontevan tarkastelukulman aineistoon. Peruslähtökohdaksi voidaan valita esimerkiksi ylin jaottelun taso tai kaikkein alin taso. Kuitenkin on toistaiseksi avoin kysymys millä hierarkian tasolla, joko päätasoilla tai niiden alatasoilla, aineiston osia on lopulta mielekästä tarkastella. Empiirisesti asiaa voisi lähestyä pyrkien mittaamaan kunkin ryhmän keskustelun *sisäistä koherenssia* tai sisäistä variaatiota verrattuna koherenssiin tai variaatioon laajemmalla tasolla. Voidaan myös kysyä, voimmeko *tunnistaa* milloin voidaan puhua esimerkiksi

keskenään keskustelevasta yhteisöstä (vrt. 1.4 Suomi24 yhteisöinä), jossa käyttäjät muodostavat kuvaa toinen toisistaan keskustelijoina, rakentavat pidempiaikaista suhdetta toisiinsa, tai jossa he muodostavat paikallisia keskustelun käytäntöjä ja keskustelukulttuuria. On varsin mahdollista, että tämä keskusteluyhteisö toteutuu pikemminkin hierarkian alimmilla tasoilla. Toisaalta on myös mahdollista, että yhteisöt syntyvät eri aihealueilla hierarkian eri tasoilla. Kysymys on mahdollinen jatkotutkimuksen aihe. Tässä osiossa tarkastelemme keskusteluja niiden päätason jaottelun mukaisina kokonaisuuksina.

Keskustelualueita on päätasolla 29, joista muutama on pikemminkin palvelun sisäisiä teknisiä alueita kuin varsinaisia keskustelualueita. Kaiken kaikkiaan pää- ja alatasoja on 2,434 (mikäli seksi-alue olisi aineistossa mukana, Allerin mukaan alueiden kokonaismäärä nousisi yli kolmeen tuhanteen). Viidennellä hierarkiatasolla olevista aihepiireistä kommenttimäärältään suurin (545,011 kommenttia) on Yhteiskunta > Uskonnot ja uskomukset > Kristinusko > Lestadiolaisuus > Vanhoillislestadiolaisuus. Kuudennelle alatasolle ulottuu esimerkiksi 34,338:lla kommentilla ”Tiede ja teknologia > Kodintekniikka > Mobiililaitteet > Kännykät > Liittymät ja palvelut > Muut liittymät”.

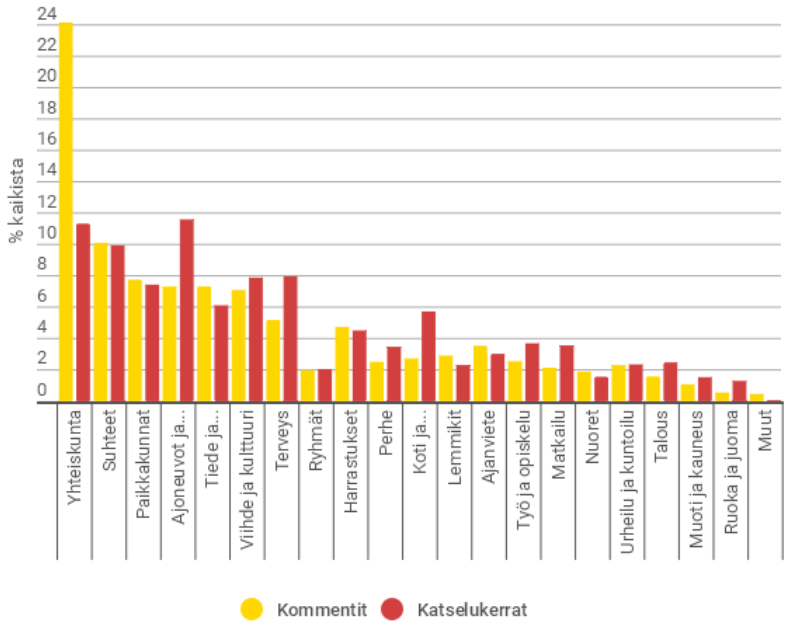
Kuinka keskustelu sitten jakaantuu määrällisesti eri aihepiireihin? Tätä on tarkasteltu päätason keskustelualueiden kautta, kommenttien, sanojen ja ketjujen avulla.



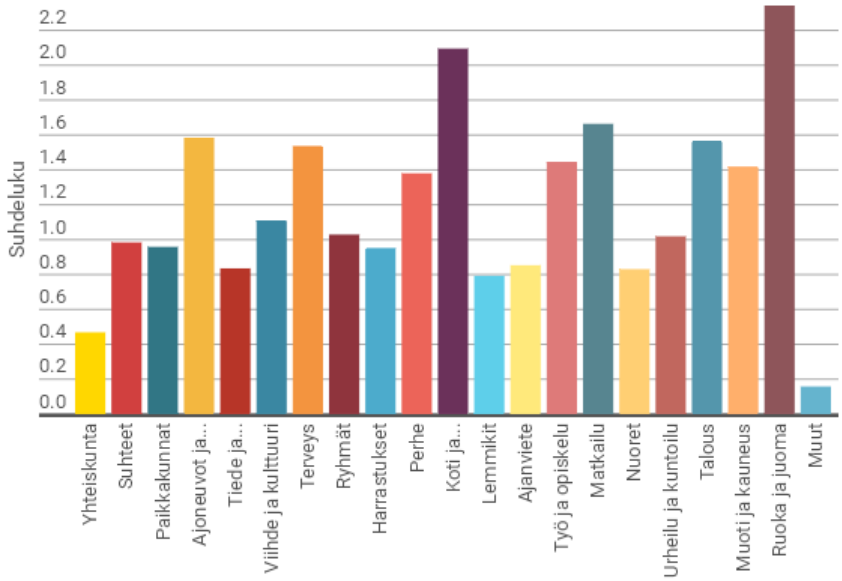
Kuva 7. Kommenttien, sanojen ja ketjujen osuudet aihepiireittäin

Kuvassa 7 yhdistetty kategoria ”Muut” sisältää kahdeksan harvinaisinta aihealuetta: Tori, Suomi24, Luotsi, Kommentaattori, Suomi24 Kehitysversio, Kumppanit, Suomi24yritys, Mainpage, Forumfrontpage. Suurin keskustelualue on Yhteiskunta, mitä selittää se että kaksi suosittua tematiikkaa, politiikka ja uskonto, molemmat sijoittuvat sen alle. Suhteet-alue käsittelee ihmissuhteita, Paikkakunnat taas Suomen paikkakuntia.

Kun tarkastellaan aineiston näyttökertoja eri keskustelualueilla, päädytään melko toisen näköiseen kuvaan (Kuva 8). Näyttökerroissa tarkasteltuna suosituimmat keskustelualueet ovat nyt 1. *Ajoneuvot ja Liikenne*, 2. *Yhteiskunta*, 3. *Suhteet*, jaetulla 4. sijalla löytyvät *Viihde ja kulttuuri* ja *Terveys*, ja 6. sijalla *Koti ja rakentaminen*.



Kuva 8. Ketjujen suosio kommenttimäärien ja näyttökertojen perusteella



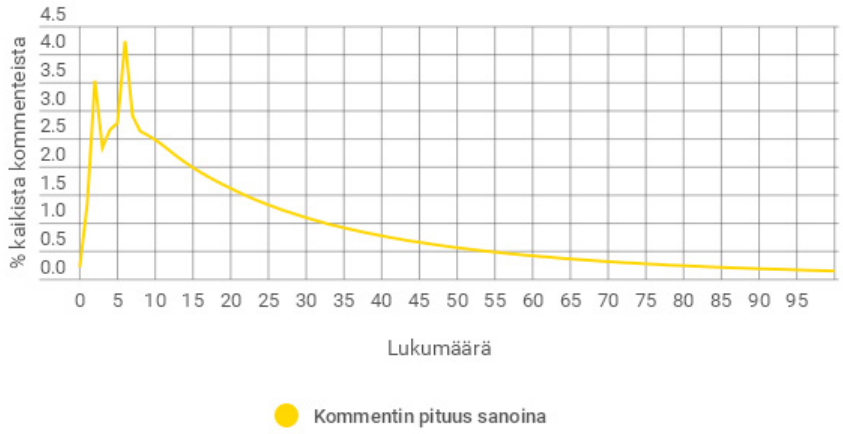
Kuva 9. Kommenttien ja näyttökertojen prosenttiosuuksien välinen suhde

Kuva 9 auttaa hahmottamaan, millä keskustelualueilla ilmenee suurin epäsuhta kirjoitusmäärien ja näyttökertojen välillä. Esimerkiksi käyttäjät lataavat Ruoka ja juoma -alafoorumien keskusteluja keskimäärin noin viisinkertaisesti verrattuna Yhteiskunta-alafoorumien keskusteluihin.

Keskustelualueiden hierarkkinen jakautuminen pääalueisiin ja alempiin alueisiin muodostaa kiinnostavan kokonaiskuvan foorumin sisäisestä elämästä. Tämä kokonaiskuva riippuu kuitenkin hyvin paljon alueiden alajaoista, ja siitä mitkä ovat alifoorumit joilla kirjoitetaan eniten. Esimerkiksi Yhteiskunta-foorumin sisältä löytyvät sekä politiikka- että uskonto-aiheiset alafoorumit, joista molempiin osallistutaan aktiivisesti, ja joiden keskustelut näyttäytyvät hyvin erilaisina monesta eri tarkastelukulmasta. Useissa analyyseissä saattaisikin olla hyödyllisempää tarkastella Yhteiskunta-foorumia sen ensimmäisen alajaon mukaisissa osissa mieluummin kuin päätasolla.

3.3 KOMMENTTIEN JA KESKUSTELUKETJUJEN PITUUSJAKAUMIA

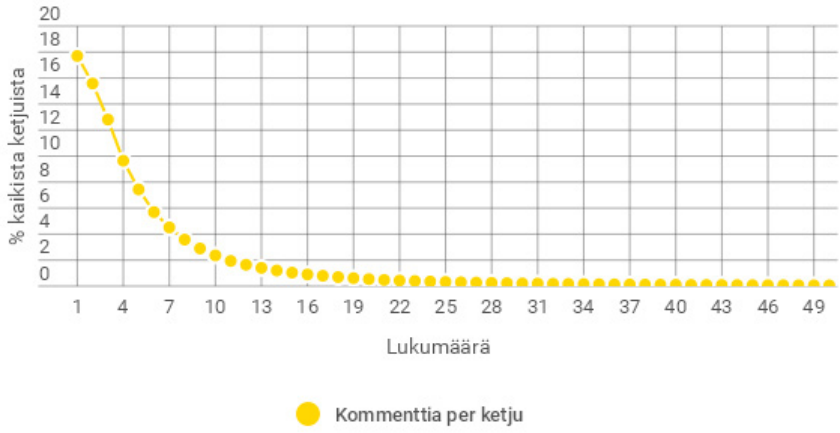
Kommenttien pituuksia on mielekästä tarkastella sanamäärinä. Kuva 10 kertoo kuinka suuri prosenttiosuus foorumin kommenteista on yhden, kahden, kolmen jne. sanan mittaisia, aina sataan sanaan saakka.



Kuva 10. Kommenttien pituuksien jakauma koko aineistossa (100 sanaan saakka)

Kuvassa 10 näkyvät erikoiset piikit ovat kahden ja kuuden sanan mittaisten kommenttien kohdalla. Näistä ensimmäisen piikin syynä ovat HTML-linkit jotka datan puhdistusalgoritmin toteutuksesta johtuen tulkittiin kahden sanan mittaisiksi. Jälkimmäisen piikin syynä ovat moderaattorien poistamat viestit, joiden tilalle sijoitetaan automaattisesti kuuden sanan mittainen vakiofraasi ”This message has been removed by ”. Tämä esimerkki valottaa osaltaan erilaisten jakaumien merkitystä aineiston tarkastelun tukena. Niiden avulla voidaan identifioida yllättäviä ilmiöitä, joihin tarttuminen ja tarkempi tarkastelu paljastaa aineiston syntymekanismeja tai sen myöhempään käsittelyyn liittyviä piirteitä, jopa ongelmia, jotka tulisi korjata. Tämä myös selventää valintoja, joita esikäsittelyssä joudutaan tekemään: pitäisikö kommentin poistosta kertova lause laskea osaksi aineistoa, vai tulisiko ne poistaa osana esikäsittelyprosessia? Jokainen tämän kaltainen yksittäinen valinta vaikuttaa kaikkien myöhempien sanafrekvenssitietojen ja muiden mittojen laskemiseen.

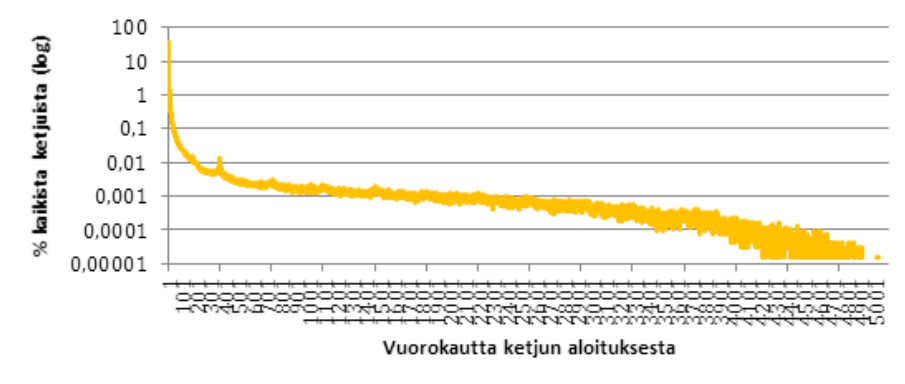
Keskusteluketjujen pituuksia voidaan tarkastella mielekkäästi joko *kommenttien määränä* keskusteluketjussa tai ketjun *ajallisena kestona*. Näistä kenties luontevampi merkitys ketjun ”pituudelle” olisi sen ajallinen kesto. Kuitenkin aktiivisen ajallisen keston määrittäminen on monessa tapauksessa vaikeaa: ketju saattaa olla aktiivinen joitain päiviä tai viikkoja, sitten painua unholaan, kunnes vuosia myöhemmin joku hakukoneen kautta tuleva kävijä lisää kommenttinsa, ja ketju alkaa taas kerätä kommentteja muilta foorumin keskustelijoilta. Onko tällainen kaksihuippuinen ketju siis vuosien mittainen keskustelu?



Kuva 11. Keskusteluketjujen pituus: kommenttimäärien suhteellinen jakauma sanoina koko aineistossa

Kommenttien määrän perusteella keskusteluketjun pituuden määrittely on yksiselitteistä. Kuva 11 näyttää kuinka suuri osa keskusteluketjuista sisältää tietyn osuuden kommenteista. Kuvaajasta voidaan nähdä esimerkiksi, että noin 12% keskusteluketjuista on saanut osakseen kolme kommenttia. Näin laskettuna keskusteluketjujen pituuksien jakauma noudattelee odotetusti nk. Zipfin lakia[24], eli painopiste on lyhyissä ketjuissa, ja toisaalta pieni määrä ketjuja on kommenttimääriltään hyvin pitkiä. Johtopäätös on, että Suomi24-foorumilla harvat keskustelut saavat pidempiaikaista huomiota, useimmat päättyvät korkeintaan muutamien kommenttien jälkeen.

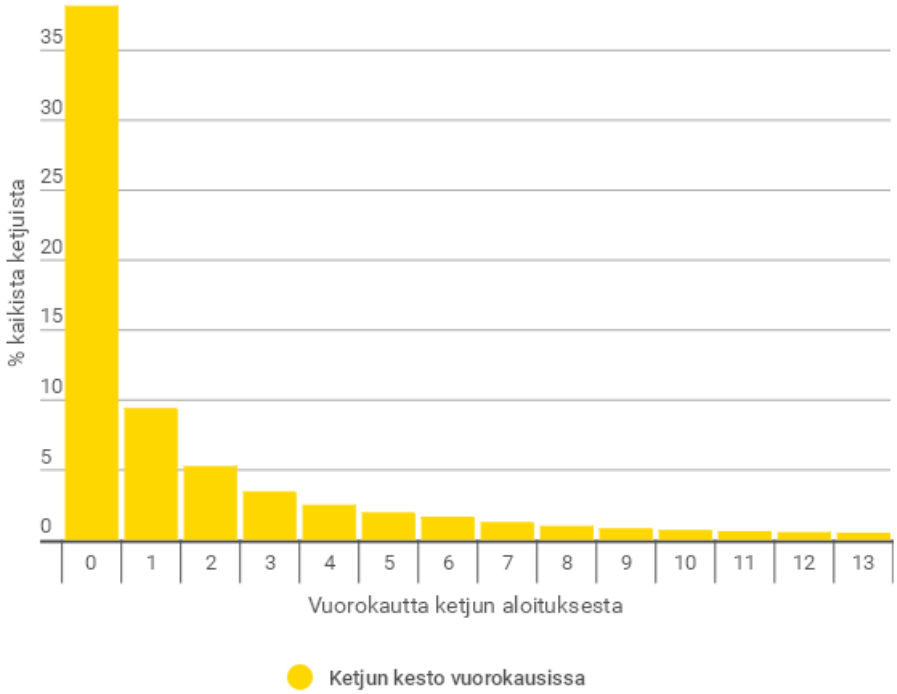
Kuva 12 esittää, kuinka suuri osa keskusteluketjuista on ollut käynnissä tietyn määrän vuorokausia. Kuvasta 12 on pääteltävissä, että foorumilla on hyvin pieni määrä ketjuja, jotka ovat olleet käynnissä jopa 14 vuotta.



Kuva 12. Keskusteluketjujen keston jakauma (logaritminen) koko aineistossa

24 https://fi.wikipedia.org/wiki/Zipfin_laki

Koska keskustelujen varhainen vaihe eli Kuvan 12 jakauman alkupää saattaisi olla määräävä tekijä sille, kuinka pitkäkestoiseksi keskustelu muodostuu esimerkiksi sen kautta, miten yksittäisen keskustelun dynamiikka lähtee käyntiin, tarkastelimme lähemmin ensimmäisten kahden viikon eli 14 vuorokauden aikana päättyvien ketjujen osuuksia koko aineistossa (Kuva 13).

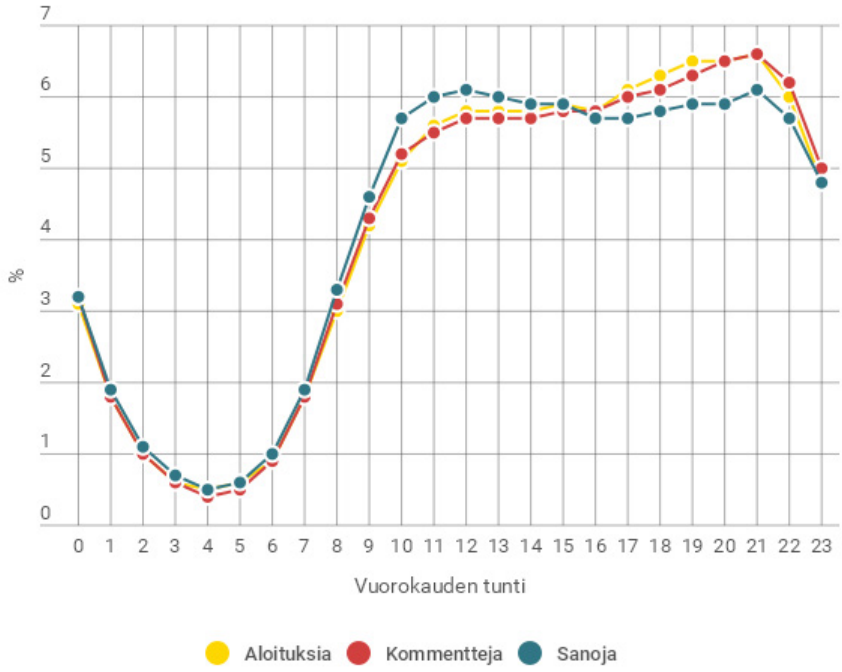


Kuva 13. Keskusteluketjujen keston jakauma, kun tarkastellaan korkeintaan 14 vuorokautta kestäneitä ketjuja

Kuva 13 kertoo kuinka suuri osa korkeintaan 14 vuorokautta käynnissä olleista keskusteluketjuista on ollut käynnissä tietyn määrän vuorokausia. Ensimmäisen 14 vuorokauden aikana päättyviä ketjuja oli koko aineistossa yhteensä 68 %. Kuvasta nähdään myös, että aloitetuista keskusteluketjuista peräti 38 % päättyy jo ensimmäisen vuorokauden aikana, ja toisen vuorokauden aikana näistä päättyy vielä noin 10 % alkuperäisestä. Selkeästi suurin osa keskusteluista päättyy siis varsin nopeasti, muutaman päivän sisällä.

3.4. KESKUSTELUN RYTMIT: AINEISTON MÄÄRÄT ERI VUOROKAUDENAIKOINA

Suomi24-foorumin kaltaisessa ympäri vuorokauden auki olevassa keskusteluyhteisössä on epäiltävissä, että eri vuorokauden aikoina ja eri viikonpäivinä keskustellaan erilaisella intensiteetillä. Jatkotutkimuksia silmällä pitäen päätimme tarkastella tätä ilmiötä lähemmin.



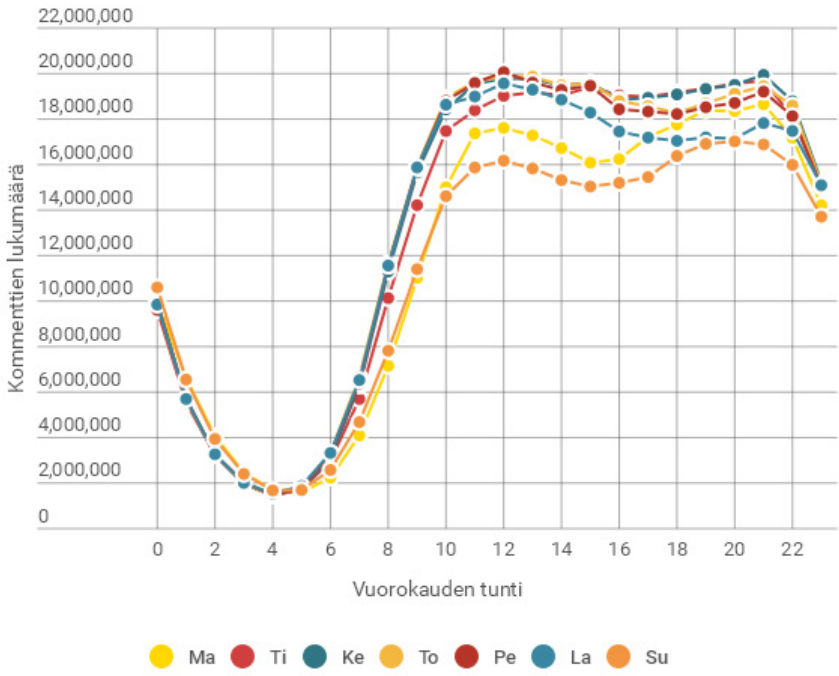
Kuva 14. Ketjujen aloitusten, kommenttien ja sanojen prosenttiosuudet kunakin vuorokauden tuntina koko aineistossa

Kuvan 14 esittämässä vuorokausirytmissä voidaan huomata seuraavat ilmiöt:

- *Lounasajan aktiivisuuspiikki:* Klo 11 – 12 esiintyy sekä pitkiä kommentteja että lukumäärällisesti paljon kommentteja. Eniten sanoja kirjoitetaan lounaalla.
- *Illan nopeat keskustelut:* Klo 21 – 23 ilmaantuu eniten uusien ketjujen aloituksia ja kommentteja, ja toisaalta silloin kirjoitetaan lyhimmät kommentit (keskiarvo = 33 sanaa)
- *Suden hetken vuodatukset:* Klo 04–05 yöllä kirjoitetaan pisimmät kommentit (keskiarvo = 41 sanaa)

- *Nukutaan:* Lukumäärällisesti vähiten kommentteja kirjoitetaan klo 05–06 välisenä aikana.

Eri viikonpäivien väliset erot tunneittain mitatussa vuorokausirytmissä käyvät ilmi alla olevasta Kuvasta 15:



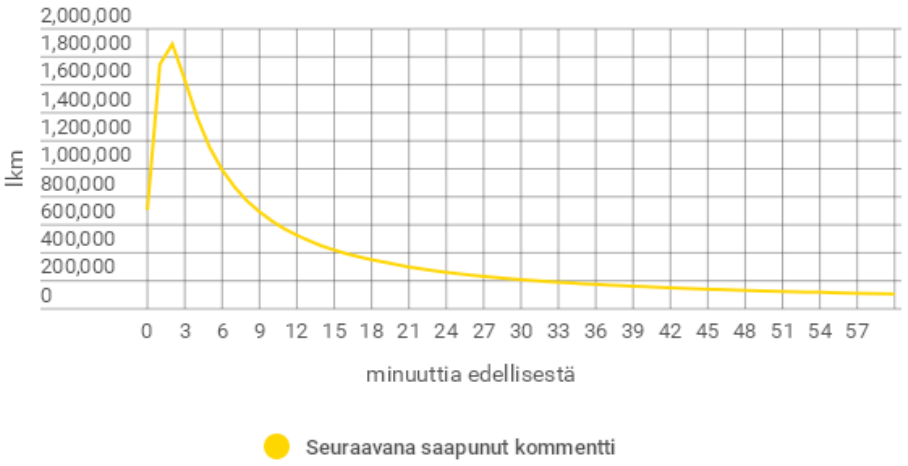
Kuva 15. Päiväkohtaiset kommentoinnin vuorokausirytmit koko aineistossa

Kuvasta 15 on tulkittavissa lauantai-iltapäivän aktiviteetin lasku. Sunnuntai pidettäneen lepopäivänä myös sosiaalisesta mediasta, ja maanantain kuluessa tapahtuu ”paluu arkeen” vähittäisenä aktiviteetin nousuna. Avoimeksi jatkokysymykseksi jää, vaikuttaako vuorokaudenaika tai viikonpäivä foorumilla käytävien keskustelujen sisältöön.

3.5 KESKUSTELUN DYNAMIIKKA

Vuorokaudenajan ja kirjoitusviikonpäivän lisäksi ihmisten välisen vuorovaikutuksen ymmärtämisen kannalta myös muut keskusteluketjujen ajalliset ominaisuudet

ovat kiinnostavia. Voidaan kysyä esimerkiksi, kuinka tiheitä tai tiiviitä keskustelut ajallisesti ovat, ja mitä muuta niissä ajallisesti tapahtuu.



Kuva 16. Kommenttien lähettämishetken jakauma ketjuissa

Kuvasta 16 voidaan havaita että noin 700,000 kommenttia on lähetetty alle minuutin kuluessa saman ketjun edeltävästä kommentista. Jakauman huippu on kahden minuutin kohdalla, eli tyypillisimmin uusi kommentti lähetetään noin kahden minuutin päästä ketjun edellisestä kommentista. Kuvaajasta on mahdollista laskea, että yhteensä 4,3 miljoonaa kommenttia on kirjoitettu alle kolmen minuutin kuluessa ketjun edeltävästä kommentista. Merkittävä osa keskustelusta vaikuttaisi siis olevan luonteeltaan nopeatahtista vuorovaikutusta foorumilla yhtä aikaa läsnä olevien käyttäjien välillä. Ilmiön syntyyn voi osaltaan vaikuttaa palvelun käyttöliittymän tapa nostaa esiin kullakin hetkellä aktiivisimmat keskusteluketjut. Esimerkki nopeatahtisesta vuorovaikutuksesta voisi olla vaikkapa televisio-ohjelmassa parhaillaan käynnissä olevasta ajankohtaisaiheesta keskusteleminen.

Jakauman pitkä häntä voi johtua puolestaan esimerkiksi siitä, että Suomi24:ssä on vakiokäyttäjiä, joilla on omat vakioalueet, joissa he käyvät hidastempoisempaa, epäsynkronista keskustelua, kukin itselleen sopivimpana vuorokaudenaikana. Toisaalta on tiedossa, että kävijät päätyvät keskusteluihin Google-hakujen seurauksena, jolloin vanhat ketjut voivat aktivoitua pidemmänkin tauon jälkeen. Tällainen ilmiö lisää tiheysjakaumaan pitkän ajallisen etäisyyden tapauksia. Kuten kommenttien syntyaikojen tarkastelussa, on mahdollista, että erilaisella intensiteetillä syntyvät keskustelut muodostuvat luonteeltaan erilaiseksi, tai että eri alueilla käydään erilaisen intensiteetin keskusteluja. Hitaamman, epäsynkronisen keskustelun esimerkki-

tapaus voisi olla vaikkapa Kuvassa 2 esitetty asiapitoinen kysymys, johon voi tulla vastaus vasta pitkänkin ajan päästä aiheesta tietävän sattuessa paikalle.

3.6 LAADULLINEN NÄKÖKULMA: MISTÄ KIRJOITETAAN JA MITÄ LUETAAN?

Eniten sanoja sisältäneet 24 ketjua on listattu alla Taulukossa 2. Siinä esiintyy mm. useita kertoja eri ajankohtina, mutta samalla otsikolla aloitettu ketju ”Kuolema ja Suru”, sekä muitakin suruaiheisia ketjuja. Myös rakkaus ja uskonnolliset teemat esiintyvät useiden ketjujen otsikossa. Lisäksi mukana vaikuttaa olevan arkisia tai leikkisiä ketjuja. Yksinäisyys on mainittu kahdessa ketjussa, ”Yksin pihalla” ja ”Yksin kotona vai hoidossa”, molemmilla kerroilla ilmeisesti liittyen lastenhoitoon ja arjen järjestelyihin. Otsikko ”jos vaihtaisit alaa” taas pohtii isoja elämänvalintoja. Aiheet, joista ihmiset haluavat kirjoittaa, piirtävät hyvin inhimillisen kuvan Suomi24-foorumin käyttäjistä oman elämänsä kannalta merkityksellisistä teemoista keskustelijoina.

Taulukko 2. Eniten *sanoja* sisältäneet keskusteluketjut

Kommentteja	Sanoja	Sanaa/ kommentti	Katsottu	Aloitettu	Ketjun otsikko
164	68 066	415	3 090	15.10.2005	Tuu mukaa- TEHÄÄ ENKAT
502	68 015	135	6 198	6.9.2010	From Your ”Leonard” To My Dearest
434	62 380	144	8 247	28.8.2010	Nyt teit sen viimeisen kerran.
820	53 866	66	21 228	8.4.2009	RAKKAUS ON KUOLEMAA VÄKEVÄMPI
167	45 928	275	1 160	26.8.2004	Kissa tippu taas pöydältä ;-)
299	45 774	153	4 490	16.9.2007	Sinä..
39	39 272	1 007	1 636	5.6.2007	301 Kysymystä!!
495	38 755	78	17 279	25.10.2008	MIKÄ ON AUTTANUT SURUSSA?
501	38 387	77	2 815	23.3.2011	Hep
501	37 473	75	1 214	19.12.2012	”Kaleva oli oma vikamme”
505	33 863	67	9 017	13.4.2009	Aikuistuminen
30	33 670	1 122	2 938	19.1.2009	Gallupit ja kyselyt
5	32 774	6 555	486	13.10.2005	Albiinotummaihoiset
503	32 725	65	9 552	27.12.2010	Kuolema ja suru
202	32 619	161	3 028	26.7.2003	Jumalan Voima
509	32 336	64	8 279	8.9.2009	SURUA JA RAKKAUTTA
501	31 819	64	9 000	2.7.2010	Kuolema ja suru
503	31 755	63	10 063	8.1.2010	Kuolema ja suru
516	31 469	61	63 772	16.12.2008	Muisteluja traktoreista
427	31 331	73	7 672	5.1.2011	To My REAL REAL AHNGELINA
502	31 099	62	3 312	5.7.2011	Kuolema ja suru

108	30 971	287	1 025	2.2.2013	Gallupit ja kyselyt
501	30 801	61	2 319	3.3.2014	Tukholman profetia nyt ajankohtainen
117	30 373	260	1 842	12.8.2004	Lutherilaisuus tarvitsee oikaisua!

Eniten kommentteja (kommenttien lukumääräkentän perusteella poimittuna) aineiston kattamana aikana saaneet keskustelut on listattu alla Taulukossa 3. Koska ketjujen kommenttimäärien maksimi on Suomi24-palvelussa 500, ja listattujen ketjujen kommenttimäärät liikkuvat tuhansissa, listaus tuli poimineeksi raskaimmin moderoidut ketjut. Tähän viitteen antaa myös suhdeluku ”sanaa/kommentti”, joka on hyvin alhainen verrattuna esimerkiksi Taulukkoon 2. Keskusteluketjuja lukemalla asia varmistuu: Esimerkiksi ”Conn Iggulden”-ketjussa[25] aloitusviestiä lukuun ottamatta kaikki 1148 viestiä on poistettu sääntöjen vastaisena, ja ketju on lisäksi suljettu.

Kun asiaa tiedusteltiin Allerilta, selvisi että osassa tapauksia kyse on ollut ns. spämmibotista (automaattisesti roskaposteja kirjoittava ohjelma, nk. spambot[26]), joka on postannut ketjun täyteen. Näistä on ollut Suomi24-palvelulle erityisen paljon vaivaa vuosina 2009 – 2010, johtaen joissain tapauksissa jopa rikosilmoituksiin spämmääjää kohtaan.

Taulukko 3. Eniten kommentteja sisältäneet keskusteluketjut

Kommentteja	Sanoja	Sanaa/ komm.	Katsottu	Aloitettu	Ketjun otsikko
5 776	3 035	1	17 504	5.6.2009	Nainen käyttää miesten vessaa
3 634	2 074	1	10 313	17.5.2009	Reputtanut
2 788	3 654	1	7 477	22.5.2009	Miten miehen saa tekee kotitöitä?
2 287	3 963	2	11 106	8.5.2009	yksin pihalla?
2 159	10 550	5	22 443	14.5.2009	Trampoliini rivitalohuoneiston pienellä pihalla?!
1 655	2 235	1	19 870	24.4.2009	Työhaastattelu ei suju
1 640	2 746	2	12 691	29.5.2009	Ongelma nuori jo 9-vuotiaana???
1 622	6 240	4	16 432	26.5.2009	miehen puhumattomuus
1 589	176	0	3 621	4.5.2007	Hi-Tech picc lite: Object file
1 571	168	0	2 683	27.5.2009	tietojärjestelmää
1 536	5 645	4	49 705	25.5.2009	Inkinen löytyi hukkuneena
1 504	878	1	2 583	25.5.2009	Yksin kotona vai hoidossa
1 300	5 769	4	968	18.7.2013	Monikulttuurisuus epäonnistunut täysin Saksassa
1 214	20	0	1 813	2.7.2009	Ein Kätzchen
1 174	2 956	3	8 249	6.5.2009	SFC-alueiden
1 148	20	0	4 655	30.9.2005	Conn Iggulden

25 http://keskustelu.suomi24.fi/t/2127674/conn-iggulden

26 https://en.wikipedia.org/wiki/Spambot

1 142	1 105	1	6 007	18.5.2009	jos vaihtaisit alaa
1 135	6 204	5	1 153	1.9.2013	Jo 40% nuorista vastustaa maahanmuuttoa
1 062	28 000	26	403	17.7.2002	Palaa hihat..
1 058	19 696	19	5 704	14.8.2009	Tervehdys kaikille!
1 048	4 658	4	10 382	26.6.2007	Prince, kaikkien aikojen paras!
1 032	4 752	5	556	6.8.2013	Islamofobian verkosto
1 017	46	0	2 577	14.8.2006	hopeanuoli dvd?
1 010	1 499	1	31 282	11.5.2009	Miten pyydän miestä ulos?
1 000	15 418	15	4 893	10.7.2009	mahdoton on mahdotonta
936	105	0	784	16.4.2013	Porvoo
910	9 382	10	40 119	26.6.2009	Michael Jackson kuollut
873	9 521	11	31 677	24.10.2010	Efexor Depot 75mg(kapseli) LOPETUS
858	2 337	3	16 813	29.4.2006	Muutto Usaan?

Tarkasteltaessa mitä ketjuja on *luettu eniten*, syntyy jälleen erilainen läpileikkauskuva foorumin käytöstä. Taulukossa 4 on koko aineistosta poimittu viestiketjut, joissa näyttökertojen määrä on suurin.

Taulukko 4. Eniten katsotut keskusteluketjut

Kommentteja	Sanoja	Sanaa/ komm.	Katsottu	Aloitettu	Ketjun otsikko
260	3 180	12	1 263 334	15.12.2006	Suomen rohkein ritari
52	689	13	1 088 206	15.5.2004	Sonera teki KALLIIN virheen minulle
569	5 810	10	848 958	29.2.2008	MANDI LAMPI KUOLLUT
150	1 189	8	612 934	19.6.2004	DC++ public hublist
213	3 882	18	587 830	13.9.2005	PERTUN PARHAAT LAUKAISUT TÄNNE!!!
336	11 765	35	542 082	22.6.2004	Cipralex?
560	5 924	11	533 327	20.3.2008	Mia Permanto
503	20 841	41	492 255	3.8.2007	AMWAY - tienaako?
269	3 903	15	466 255	18.11.2007	WinCapita
76	1 507	20	461 482	7.5.2004	On hyvä, että Suomi
513	9 746	19	459 611	2.11.2006	Renkaiden ilmanvaihto
230	3 787	16	449 799	15.1.2004	HUIJAUS Idols-finaalissa
255	5 383	21	443 262	18.2.2005	Omatekoinen kebab
134	5 391	40	430 698	20.6.2005	veritulppa!
670	5 843	9	415 784	2.5.2007	Susan Kuronen alasti HYMYSSÄ
41	261	6	397 716	30.8.2005	Big Brother Alastonkuvat!!
404	5 226	13	393 595	22.4.2004	Rehellinen gallup: Paljoko tienaat?
170	1 731	10	390 729	26.5.2004	1.6. EI TANKATA
43	300	7	372 257	17.2.2007	ILMAINEN NUMEROHAKU
419	3 670	9	368 034	12.4.2007	Nettihäiriköt piinaavat Annaa
97	2 388	25	347 652	16.3.2004	Apteekin oma raskaustesti
484	15 221	31	344 998	5.11.2005	PIXMANIA.
330	3 662	11	330 496	26.8.2004	Nykästä epäillään tapon yrityksestä
165	3 951	24	319 870	15.5.2004	peräaukon kutina

302	2 946	10	310 252	18.5.2008	Ravintolan edessä ammuskeltu
202	6 186	31	305 056	2.9.2003	Automyyjien suusta
19	404	21	301 419	30.10.2005	jouluruno tai värssy
262	11 744	45	293 223	14.1.2005	Tietzen oireyhtymä
109	607	6	287 696	25.4.2006	äitienpäivä runo
70	901	13	285 117	25.9.2004	ihana omenapiirakka
364	9 267	25	283 839	9.6.2006	Lyrica
66	1 106	17	283 701	1.7.2004	halvin talopaketti
256	1 763	7	282 707	22.5.2006	Awa nainen?
92	721	8	282 704	28.9.2004	Jasmin Mäntylä nakuna Hymyssä!

Näytetyimpien ketjujen perspektiivistä katsottuna kävijöitä kiinnostaa eniten *terveys* (Cipralex, veritulppa, Apteekin raskaustesti, peräaukon kutina, Tiezen oireyhtymä), *julkkikset* (Mandi Lampi, Mia Permanto, Susan Kuronen, Nykänen, Jasmin Mäntylä), *alastomuus* (alasti Hymyssä, alastonkuvat), *raha* (paljonko tienaat, halvin talopaketti, Sonera teki kalliin virheen, ilmainen numerohaku, tienaako), *huijaukset* (WinCapita, Idols-huijaus), *väkivalta* (Ravintolan edessä ammuskeltu, tapon yritys) ja *arkiset asiat* (renkaiden ilmanvaihto, 1.6. ei tankata, ihana omenapiirakka, jouluruno, äitienpäiväruno, omatekoinen kebab).

Edellä kuvatut esimerkit olivat suoraan aineistosta löytyviin metatietoihin kuten aikaleimoihin perustuvia tapoja järjestää aineistoa ja tutkia keskusteluja. Tuloksista voidaan havaita että nämä matemaattisesti ja teknisesti sinällään hyvin yksinkertaiset tarkastelukulmat jo antavat toisistaan keskenään poikkeavan kuvan Suomi24-aineistosta sekä sen käyttötavoista ja merkityksistä niin lukijoille kuin kirjoittajille. Tarkastelun syventäminen vaikkapa viemällä näinkin yksinkertaisia tarkasteluja yksittäisten alifoorumien tasolle voisi avata keskenään erilaisia keskustelukulttuureja kiinnostavalla tavalla. Toisaalta se rohkaisee tulevaisuudessa eri keskustelualueisiin perehtyviä tutkijoita keskustelemaan ”oman” alueensa ominaispiirteistä muiden tutkijoiden kanssa ja mahdollisesti löytyvien keskustelujen erojen kautta ymmärtämään ”omalla” alueella käytyjä keskusteluja aiempaa syvällisemmin.

3.7 TIETOA KESKUSTELIJOISTA

Aineistossa ei ole mukana tietoa aineiston lukijoista (lukuun ottamatta ketjujen näyttökertoja), ainoastaan foorumille kirjoittaneista. Kuten edellä on kuvattu, kirjoittajat jakautuvat rekisteröityneisiin ja rekisteröitymättömiin käyttäjiin. Rekisteröityneitä käyttäjiä on ajanjaksolla ollut yhteensä 59,612 kpl. Heidän kirjoittamansa osuus kommenteista koko aineistossa on 7,2 %. Osuus on vaihdellut vuosittain ja viime vuosina ollut 8 %:n molemmin puolin, kuten käy ilmi Kuvasta 17:

Kuva 17. Rekisteröityneiden kirjoittajien osuus aineistossa vuosittain

Käyttäjien aktiviteetin tutkiminen yksittäisten kävijöiden eli käytännössä heidän kirjoittajanimimerkkiensä kautta on aihe, joka luonnollisesti kiinnostaa monia. Tarkastelun hankaluudesta kertoo kuitenkin osaltaan alla oleva Taulukko 5, jossa näkyy aineiston 30 yleisintä nimimerkkiä.

Taulukko 5. Kolmekymmentä useimmiten käytettyä nimimerkkiä

Lukumäärä	Nimimerkki
236315	...
77083	Nimetön
74282	ap.
73852	....
69059	.
51535	vapunen
51259	\
50873	.....
47230	M-Kar
45970	???
40869	.....
35802	-
34570	minä
34086	ap
33401	---
32793	Kössönöm
31671	RepeRuutikallo
29038	Ratikkakuski
28334	.....
27424	Mover
26373	Sähköteurastaja
26270	kuunteleva_kirkko

25768	?
25373	Tosi on
23198	.....
22833	hunksz
22620	TeuvoSuni
20947	.....
20933	j.jussiW-T
20713	Kollimaattori

Nimimerkeistä osa on rekisteröimättömiä (esim. ”…”, ”Nimetön” ja ”---”). Rekisteröidyistä nimimerkeistä ”kuunteleva_kirkko” on Suomi24:ssä aktiivinen yhteisökäyttäjä, samaten rekisteröityneitä käyttäjiä ovat esim. Kössönöm, RepeRuutikallio ja Ratikkakuski.

Olisi kenties houkuttelevaa tutkia pelkästään rekisteröityjen kirjoittajien erilaisia profiileja ja palvelunkäyttötapoja. Vaikka rekisteröityneiden käyttäjien määrä sinänsä on suuri (59 612), käyttäjien tekstejä lukiessa törmätään aineiston anonyymiyttä koskeviin eettisiin kysymyksiin, huolimatta siitä, että nimimerkit ovat periaatteessa anonyymejä, eikä data yleiseltä rakenteeltaan sinänsä identifioi käyttäjiä. Tutkittavilla on yleisesti noudatettujen tutkimuseettisten periaatteiden mukaisesti oikeus tietää olevansa tutkimuksen kohteena. Mitä yksityiskohtaisemmin yksittäisten ihmisten henkilökohtaiseen elämään, kokemuksiin ja tuntemuksiin paneudutaan, sitä tärkeämpää tämä on huomioida. Käytännössä etenkin Suomi24-vakiokirjoittajat paljastavat vuosien varrella intiimejä asioita itsestään. Foorumin sisältöjä lukiessa selviää, että jotkut kirjoittajista on identifioitu yhteisön keskuudessa jopa kotipaikkakuntaa ja internet-sivuja myöten. Myös tutkijalle herää jo keskustelunimimerkkejä tarkastellessa kysymys, onko nimimerkki ”EtunimiSukunimi” itse asiassa todellinen henkilö nimeltä ”Etunimi Sukunimi”.

Suomi24-aineistot on tuotettu julkiselle foorumille, mutta käytännössä keskusteluja käydään usein varsin rajatuissa, tiettyjen teemojen ympärille järjestyneissä yhteisöissä. Varautuneisuus ja harkinta sekä tutkimuseettisten tarkastelukulmien soveltaminen on erityisen aiheellista, kun tutkitaan herkästi haavoittuvia ihmisryhmiä. Näitä epäilemättä löytyy myös Suomi24-kirjoittajien joukosta. Kirjoittaessaan foorumiin ihmiset luottavat anonymiteetin suojaan. Kun aineistoa tutkitaan, on pyrittävä löytämään tapoja, joilla yksilönsuoja ei vaarannu tutkimuksen tai tutkimusvälineistön kehittymisen myötä. Jo yksittäisen nimimerkin profilointi sen perusteella mille foorumeille tai mistä teemoista hän aktiivisimmin kirjoittaa, on omiaan luomaan tarkan kuvan ihmisen kiinnostuksen kohteista, mahdollisesti jopa äskettäisistä elämäntapahtumista ja niiden sijoittumisesta ajallisesti.

Ihmistieteellisen tutkimuksen eettisenä perustana on tutkittavien kunnioittaminen, heidän kulttuurinsa ja oikeuksiensa huomioiminen. Suomi24-tutkimuksen kannalta kulttuurinen sensitiivisyys tarkoittaa sitä, että aineistoa tulisi tulkita tut-

kittavien lähtökohdasta. Tämä ei tarkoita, että foorumien sisällöt tulisi hyväksyä sellaisenaan vaan sitä, että ymmärretään aineiston tuottamisen konteksti ja keskustelukulttuurin käytännöt. Suomi24-aineiston pohjalta tehdystä tutkimuksesta ei tulisi olla foorumin käyttäjille haittaa, vaan tutkimuksen tulisi mielellään hyödyttää myös tutkimuskohdetta. Jo hankkeen käynnistymisestä uutisoitaessa Suomi24-foorumilla syntyi hankkeesta keskustelua, ja on odotettavissa, että tutkimuksesta raportoitaessa tulokset päätyvät keskusteluyhteisöön. Onkin hyvä pohtia, tulisiko Suomi24-pohjaista tutkimusta eteenpäin vietäessä käydä keskustelua myös tutkijoiden ja foorumin aktiivikäyttäjien välillä.

4. JOHTOPÄÄTÖKSIÄ

Yhteiskuntatieteellisen, ihmistieteellisen, kielitieteellisen ja menetelmällisen tutkimuksen kannalta Suomi24-aineisto on tutkimuskohteena moni-ilmeinen. Aineisto tarjoaa poikkeuksellisen laajan ja pitkäkestoisen kokoelman arkista keskustelua ja puheenomaista kieltä. Aineisto näyttäisi siis olevan arvokas tutkimuksen resurssi ja hedelmällinen tutkimuskohde useiden eri tieteenalojen näkökulmista.

Suomen kielen tutkijoiden kannalta aineiston avulla voi tutkia kielen muutosta, kuten uudissanojen syntyä[27], ja erilaisten sananmuotojen yleisyyttä, siis ilmiöitä, jotka näkyvät ensin puhekielessä ja puhekielen omaisessa kirjoittamisessa. Aineisto avaa mahdollisuuden kielellisen vuorovaikutuksen tutkimiseen niin keskustelun rytmien ja dynamiikan, eri keskustelijoiden, kuin keskustelun sanaston ja sisältöjen kautta.

Kieliteknologian menetelmänkehityksen kannalta aineiston puhekielisyys ja suuri määrä tarjoaa aiempaan verrattuna poikkeuksellisen hyvän lähtökohdan esimerkiksi luoda puheenomaisen kielen *kielimalleja*, eli tilastollisia malleja sanojen välisistä todennäköisyyksistä tekstissä. Kielimalli on välttämätön osa vaikkapa puheentunnistusjärjestelmää.

Terveyden tai muiden erityisaiheiden tutkimuksen kannalta Suomi24-aineisto näyttäytyy mahdollisuutena tutkia kansalaiskeskusteluja tai niiden muutoksia, myös suhteessa ajankohtaisiin muihin tapahtumiin. Hyvinvoinnin tai köyhyyden tutkijoita aineistossa kiinnostaa keskustelu hyvinvointiyhteiskunnasta tai elämässä selviämisestä vaikkapa velkakurimuksessa. Vihapuhetutkijoiden näkökulmasta aineistosta on mahdollista tutkia vihapuheen muotoja, laajuutta, yleisyyttä eri teemafoorumeissa, suhteessa käyttäjäkuntaan sekä sen muutoksia ajassa. Myös rakkauspuhetta, surupuhetta, tukipuhetta, empatiapuhetta, arjen ongelmanratkaisupuhetta ja oman sielunelämän käsittelyä koskevaa puhetta aineistosta löytyy runsaasti. Useat edellä mainituista keskustelun muodoista on tunnistettu myös vertaistukitutkimuksessa vertaistuen eri muodoiksi tai funktioiksi. Voitaisiinko aineiston jotakin osaa tarkastella kymmenien tai satojen vertaistukiryhmien joukkona?

Tämän aineistokuvauksen tarkoituksena on ollut edistää tutkimustapaa, joka ottaa etäisyyttä käyttäjäkeskeiseen tutkimusotteeseen ja tarjoaa näkökulmana tilalle toisaalta tekstipohjaista ja toisaalta pikemminkin kollektiivisia ilmiöitä kuin yksilöitä tutkivaa tarkastelutapaa. Teksteihin ja niiden järjestymiseen keskittyvässä tarkastelussa tarvitaan monenlaisia menetelmiä ja välineitä, jotka avaavat ai-

27 Ulla Tuomarla, Perpanssihavaintoja, *Citizen Mindscapes Tutkimusblogi*, 21.1.2016. http://blogs.helsinki.fi/citizenmindscapes/2016/01/21/perpanssihavaintoja/

neistoa erilaisista näkökulmista, eri tasoilta ja lähtien erilaisista tiedon tarpeista. Tässä raportissa on käytetty muutamaa erilaista lähestymistapaa, joilla olemme hahmotelleet aineistoa tai tarttuneet siihen eri näkökulmista. Esiteltyjä näkökulmia voi pitää kokeilevina kurkistuksina. Pyrkimyksenämme on ollut rakentaa siltaa eri alojen sisältötutkijoiden, kielentutkijoiden ja toisaalta kieliteknologien ja data-analyytikkojen välille.

Tehtävänantoja on rakennettu lähtien sisällöllisiä tutkijoita kiinnostavista kysymyksistä, jotka ovat tarkentuneet tiiviissä vuoropuhelussa menetelmä- ja data-analyytikoiden kanssa. Tarkastelukulmina meitä ovat houkuttaneet esimerkiksi arjen rytmit, muodon ja sisällön vuoropuhelu, uudissanojen synnyn tunnistaminen, vuorovaikutuksen rytmit, laadut ja tunteet, keskustelun purskeisuus, uusien aiheiden synty ja leviäminen, sekä ympäröivän yhteiskunnan tapahtumien luomat impulssit keskusteluiden synnyttäjinä.

Sisältönäkökulmasta meiltä on toivottu vastauksia myös siihen, mitkä teemat ovat kulloinkin ajankohtaisia, miten eri aihepiirit liittyvät toisiinsa, tai miten voisimme saada selville ja visualisoituna jonkin teeman keskusteluun liittyvät keskeiset käsitteet ja jotain näiden välisistä suhteista. Keskustelujen määrän puolesta tämä aineisto todellakin tarjoaa mahdollisuudet näiden kysymysten selvittämiseen myös tilastollisen tekstianalyysin keinoin.

Tutkijoille tarjolla oleva Korp-työkalu, jonka käytöstä olemme esittäneet esimerkin, tarjoaa tutkijalle nopean ja melko helposti lähestyttävän tavan tehdä niin yksinkertaisia kuin monimutkaisempiakin hakuja Suomi24-aineistoon. Työkalun tuottamat visualisoinnit ovat kuitenkin hyvin rajallisia, eivätkä ne toistaiseksi tarjoa mahdollisuuksia edetä useisiin kiinnostaviin tutkimuskysymyksiin.

Monia mahdollisia aiheita jäi tämän raportin ulkopuolelle tai kartoittamatta. Esimerkiksi temaattisen analyysiin tai ryhmittelyn (topic modeling[28]) tai käsitteiden välisiä suhteita kartoittavaan suuntaan emme aineiston kuvauksen yhteydessä ryhtyneet, siitä huolimatta että näiden avulla voidaan selvittää monia kiinnostavia kysymyksiä. Aiheesta on kuitenkin paljon tutkimusta sekä valmiita menetelmiä tarjolla. Tässä raportissa on pyritty avaamaan vähemmän tutkittuja näkökulmia, esimerkiksi keskustelun rytmejä, jotka ehkä ovat vähemmän tavallisia tekstin louhinnan alueella, mutta nimenomaan yhteiskuntatieteilijöitä ja vaikkapa arjen elämän tai inhimillisen vuorovaikutuksen tutkijoita kiinnostavina voisivat johtaa aivan uudenlaisten ihmis- ja yhteiskuntatieteellisten data-analyysityökalujen ja menetelmien kehittelyyn.

Työkalujen jatkokehittämisen kannalta toivomme että nämä aineistoon tehdyt kurkistukset avaavat näkymää siihen, mitä humanistiset tai yhteiskuntatieteelliset kysymyksenasettelut voisivat olla, ja näin synnyttää ajatuksia koskien nykyisten

28 https://en.wikipedia.org/wiki/Topic_model

kieliteknologisten tai tekstianalyyttisten työkalujen laajentamisen mahdollisuuksia, tai johtaa kokonaan uusien työkalujen rakenteluun. Digitaalisten ihmistieteiden näkökulmasta tarvitsemme erilaisten mikroskooppien ja makroskooppien laajempaa tuotantoa. Tässä yhteydessä olemme rakentaneet omat karkeat tutkimusvälineemme työn edetessä ns. purkkaviritys[29]-periaatetta soveltaen: suurin osa tarvitsemistamme aineistoanalyyseistä on tuotettu tutkimusryhmässä nopeasti tätä varten kirjoitetuilla Perl-ohjelmilla. Kuvat on tuotettu Exceliä käyttäen.

Digitaalisten ihmistieteiden työvälineiden kehittäminen ihmis- ja yhteiskuntatieteitä paremmin palvelevaan suuntaan edellyttää aktiivista vuoropuhelua ja yhdessä työskentelyä. Olemme tätäkin tehdessä havainneet, että monipuolisemmin ilmiöitä huomioonottavia tuloksia syntyy, kun mukana on sekä aineisto- ja menetelmäosaajia että toisaalta humanisteja ja yhteiskuntatieteilijöitä. Yhteistyöllä on mahdollista löytää uusia tutkimuskysymyksiä, joihin pystytään menetelmien kehittyessä myös vastaamaan.

29 http://urbaanisanakirja.com/word/purkkaviritys/

5. TUTKIMUKSEN TYÖKALUPAKKI JA YHTEISTYÖN MUODOT

Suomi24-aineisto on avattu tutkimuskäyttöön avoimen datan hengessä. Tavoitteena on rakentaa sosiaalisen median tutkimuksen työkalupakki, johon vähitellen kerrytetään välineitä ja resursseja edesauttamaan tämän tyyppisten aineistojen tutkimusta. Tämän aineistokuvauksen tarkoituksena on helpottaa yhteistä dialogia tarjoamalla jaettua sanastoa ja ymmärrystä aineiston ominaispiirteistä. Toivomme, että aineiston käyttäjät omalta osaltaan pohtisivat, mitä he voisivat vuorostaan tarjota yhteiseen työkalupakkiin. Nämä voivat olla esimerkiksi sanastoja jostakin tietystä teemasta, aineiston annotointeja jonkun aineistonäytteen osalta, ohjelmistoja tai data-analyysin työkaluja.

Yhteistyön tavoitteena on vastavuoroisuutta ja eettisiä käytäntöjä tietoisesti edistävä tutkimuskollektiivi. Keskeneräisistä asioista toivotaan keskustelua, jotta yhteiseen toimintaan liittyvät ongelmat voidaan ratkoa. Toimivien käytäntöjen löytyminen on olennaista, jotta monitieteinen ja monimenetelmäinen tutkimuskollektiivi voi toimia. Yhteistyön pohjana tarvitaan vastavuoroisen tekemisen kulttuuria ja toimintatapoja, joilla yhteistyö muodostuu kaikkia osapuolia rikastavaksi sekä turvalliseksi. Tutkijoiden väliseen keskusteluun voi osallistua *Connect Lab*-tapaamisissa[30]. Keskustelua käydään lisäksi kaikille avoimessa Facebook-ryhmässä nimeltä *Citizen Mindscapes*. Yhteistyön muotoja, niiden kehittymistä ja onnistumisia dokumentoidaan myös tutkimusblogissa[31].

30 http://blogs.helsinki.fi/citizenmindscapes/2016/01/14/connect-lab-suomi24-aineiston-yhteisollinen-tutkiminen/

31 http://blogs.helsinki.fi/citizenmindscapes/

Kuluttajatutkimuskeskus

Tässä raportissa luodaan katsaus Suomi24-keskustelufoorumin aineistoon, joka on ollut saatavilla Kielipankissa keväästä 2015. Raportissa kuvataan aineiston syntytapaa ja kontekstia, eli Suomi24-informaatioarkkitehtuuria, aineiston määriä ja luonnetta, aineiston tallennusmuotoja, ja tapoja, joiden avulla tutkija voi aineistoon perehtyä. Lisäksi viritetään esimerkinomaisesti tutkimuksellisia näkökulmia, joihin aineistotyöskentelyllä voisi edetä.

Digitaalisten tekstiaineistojen laajamittainen yhteiskuntatieteellinen tutkiminen on vasta aluillaan. Raportti tarjoaa taustan, jota vasten laajasta aineistosta tehtyjen yksittäisten hakujen tuloksia voi suhteuttaa ja tulkita. Lisäksi aineiston käyttöesimerkkien ja erilaisten lähestymistapojen esittely voi tarjota virikkeitä aineiston hyödyntämiselle omassa tutkimuksessa.

Aineistoraportin tavoitteena on tuoda esille kvantitatiivisen ja kvalitatiivisen aineistotutkimuksen mahdollisuuksia. Toivon mukaan näihin tarttuminen voi myös edistää analyysityökalujen kehittymistä tavalla, joka palvelee entistä paremmin paitsi eri tieteenalojen tutkijoita, myös journalisteja ja aiheesta kiinnostuneita kansalaisia.

HELSINGIN YLIOPISTO
VALTIOTIETEELLINEN TIEDEKUNTA

ISBN 978-951-51-1065-7