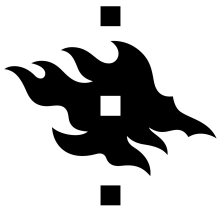


HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Pro Gradu -tutkielma
Aluetiede
Kaupunkimaantiede

Suomen- ja englanninkieliset digitaaliset kaupunkitilat Helsingissä? - Case Instagram

Tuomas Väisänen

2018

Ohjaajat: Tuuli Toivonen ja Tuomo Hiippala

HELSINGIN YLIOPISTO
MATEMAATTIS-LUONNONTIETEELLINEN TIEDEKUNTA
GEOTIETEIDEN JA MAANTIETEEN LAITOS
MAANTIEDE

PL 64 (Gustaf Hällströmin katu 2)
00014 Helsingin yliopisto

HELSINGIN YLIOPISTO − HELSINGFORS UNIVERSITET − UNIVERSITY OF HELSINKI

Tiedekunta/Osasto − Fakultet/Sektion) Faculty	Laitos − Institution) Department
Matemaattis-luonnontieteellinen tdk.	Geotieteiden ja maantieteen laitos

Tekijä − Författare) Author
Tuomas Väisänen

Työn nimi − Arbetets title) Title
Suomen- ja englanninkieliset digitaaliset kaupunkitilat Helsingissä? - Case Instagram

Oppiaine − Läroämne) Subject
Kaupunkimaantiede

Työn laji − Arbetets art) Level	Aika − Datum – Month and Year	Sivumäärä − Sidoantal – Number of Pages
Pro Gradu -tutkielma	Marraskuu 2018	169

Tiivistelmä − Referat) Abstract

Ubiikkien internet yhteyksien, mobiililaitteiden ja nousevien sosiaalisen median käyttäjämäärien myötä massadataa yksilöistä, kaupungeista, kulttuureista ja yhteiskunnista tuotetaan suuria määriä. Tarve nykyaikaisen nopeasti muuttuvan digitalisoituvan maailmamme ymmärtämiselle kasvaa jatkuvasti. Tämä tekee sosiaalisen median massadata-aineistoista arvokkaan maantieteelliselle tutkimukselle, mutta massadata on laadultaan, määrältään ja monipuolisuudeltaan sellaista, että sen käsittely vaatii uusia ja tehokkaita poikkitieteellisiä metodeja. Sosiaalisen median spatiaalisen massadatan tarkastelu mahdollistaa sen kautta välittyvien digitaalisten kaupunkitilojen tutkimisen.

Tässä työssä selvitetään Helsingin alueelta tehtyjen suomen- ja englanninkielisten Instagram-julkaisujen lingvististä sisältöä tarkoituksena (1) selvittää kyseisten Instagram-julkaisujen aiheet, (2) aiheiden spatiotemporaalinen rakenne, (3) kielten väliset erot aiheissa ja (4) lingvististen teknologioiden soveltuvuus maantieteelliseen tutkimukseen. Näiden päämäärien saavuttamiseksi, työssä yhdistellään luonnollisen kielen prosessoinnin ja geoinfromatiikan metodeja soveltamalla aihemallinnusta eri spatiaalisissa mittakaavoissa. Erilaisia visualisointitekniikoita hyödynnetään kielikohtaisten aiheiden analysointiin Helsingin eri kaupunginosissa.

Tulokset osoittavat, että suomen- ja englanninkielisten Instagram-julkaisujen aiheet eroavat toisistaan spatiaalisesti, temporaalisesti, vaikka molempien kielten osalta aineisto osoittautui sävyltään ja temporaaliselta rytmiltään samankaltaiseksi. Tuloksena syntyneet spatiaaliset aihemallit vaikuttavat tukevan teoreettista keskustelua digitaalisista kaupunkitiloista, mutta useaa sosiaalisen median alustaa ja useita poikkitieteellisiä metodeja yhdistävää tutkimusta tarvitaan vahvemman empiirisen näytön saamiseksi. Luonnollisen kielen prosessinnin metodit vaikuttavat soveltuvan hyvin maantieteelliseen tutkimukseen, erityisesti sosiaalisen median aineistojen kanssa.

With ubiquitous internet access, mobile smart devices and increasing amounts of social media users, big data about individuals, cultures and societies is being produced in vast quantities. This renders social media data valuable for geographic research, but such data requires novel interdisciplinary methods. Spatial social media big data enables the investigation of emergent digital urban spaces.

This Master's thesis analyses the linguistic content of English and Finnish Instagram posts uploaded from Helsinki, Finland, in order to identify (1) common topics, (2) how these topics evolve over space and time, (3) differences in finnish and english spatiotemporal topics and (4) applicability of natural language processing in geographic research. To do so, the study combines methods from natural language processing and geoinformatics, applying the technique of *topic modelling* at various spatial scales. Different visualization techniques are used to support the analysis of language-specific topics in different neighborhoods of Helsinki.

The results show that the topics of Finnish and English Instagram posts differ spatially, temporally and at different scales, although the data for both languages was shown to be similar in sentiment and temporal rhythm. The resulting spatial topics models appear to support theories of digital urban spaces to some extent, but more comprehensive analysis of digital urban spaces requires data from various social media platforms and diverse interdisciplinary methods. Natural language processing methods hold much potential for geographic research, particularly for analyzing social media data.

Avainsanat – Nyckelord) Keywords
kaupunkimaantiede, sosiaalinen media, massadata, luonnollisen kielen prosessointi, tekoäly

Säilytyspaikka – Förvaringställe – Where deposited
Helsingin yliopisto, Keskuskirjasto

Muita tietoja) Övriga uppgifter) Additional information

Alkusanat ja kiitokset **5**

Sanasto **6**

Kuvat ja taulukot **7**

1.0 Johdanto **8**

1.1 Työn taustoittaminen 11

1.2 Kaupunkitila 13

1.2.1 Digitaalinen kaupunkitila 15

1.2.2 Erilaiset ryhmät kaupungissa 21

1.3 Sosiaalinen massadata 24

1.3.1 Instagram ja muut sosiaaliset mediat 27

1.4 Sosiaalinen media maantieteellisessä tutkimuksessa 29

2.0 Aineisto ja menetelmät **32**

2.1 Instagram-aineisto 34

2.1.1 Aineiston kuvaus 36

2.1.1.1 Käyttäjät, aihetunnisteet ja kohdepisteiden sijainnit 36

2.1.1.2 Ajallinen rakenne 42

2.1.1.3 Alueellinen rakenne 47

2.2 Aineiston käsittely 54

2.2.1 Esikäsittely 54

2.2.1.1 Automaattinen kielentunnistus 56

Menetelmän kuvaus 56

Kielentunnistus Instagram-aineistolle 58

2.2.1.2 Kuvatekstien lemmatisointi 60

Menetelmän kuvaus 60

Lemmatisointi Instagram-aineistolle 61

2.2.2 Aihemallinnus LDA-menetelmällä 62

Menetelmän kuvaus 62

Aihemallinnus Instagram-aineistolle 66

3.0 Analyysi **69**

3.1 Kielentunnistuksen tulokset 69

3.2 Lemmatisoinnin tulokset 84

3.3 Aihemallinnuksen tulokset 86

3.3.1 Helsingin alueen tarkastelu 87

3.3.1.1 Aiheet ajassa 91

3.3.1.2 Aiheet spatiaalisesti 99

3.3.2 Valittujen kaupunginosien tarkastelu 103

3.3.2.1 Kallio 105

3.3.2.2 Töölö 115

3.3.2.3 Suomenlinna 125

3.3.2.4 Adjektiivimallinnus 134

4.0 Keskustelu **136**

4.1 Julkaisujen aiheet liittyvät vapaa-aikaan 136

4.2 Aiheiden spatio-temporaalinen rakenne on kaksijakoinen 140

4.2.1 Aiheiden erot ajassa 141

4.2.2 Aiheiden erot alueellisesti 143

4.3 Helsingin digitaalinen kaupunkitila näyttäytyy eri lailla kyseisten kieliryhmien julkaisuissa 146

4.4 Kieliteknologian menetelmät soveltuvat kaupunkitilan tutkimukseen 150

4.5 Lopuksi 154

5.0 Kirjallisuus **158**

Alkusanat ja kiitokset

Isaac Asimov kirjoitti scifi-kirjallisuuden merkkiteoksessaan, Säätiö-trilogiassa, psykohistoriasta, joka oli ihmiskunnan tulevaisuutta ennustava tieteenala. Psykohistoria ei pystynyt ennustamaan yksittäisten ihmisten tulevia päätöksiä yhtään paremmin kuin kaupungin, valtion tai kokonaisen planeetan väestön päätöksiä. Se perustui ajatukseen siitä, että tarpeeksi suurena massana kuvattuna ihmiskunnan käyttäytymisen mallintamisesta tuli luotettavaa ja siten myös ihmismassojen käyttäytyminen oli ennustettavissa. Ennustuksien toteutumisen mallin mukaan oli olennaista, ettei suuret ihmismassat saisi tietää ennusteiden olemassaolosta kokonaisuudessaan, koska muuten ihmismassat voisivat valita tekevänsä toisin, jolloin mallin ennustama lopputulos ei toteutuisi vaan vääristyisi. Säätiö-kirjasarja sijoittuu erittäin kaukaiseen tulevaisuuteen, jossa ihmiskunta on asuttanut koko galaksimme ja on käytännössä universaalin imperiumin hallitsema, jonka romahdusta ei pysty enää estämään, mutta romahdusta seuraavaa kurjuuden ja sodan runtelemaa aikakautta pystyy pienentämään useista kymmenistä vuosituhansista vain yhteen vuosituhanteen psykohistorian avulla. Asimovin kirjasarjan kauan sitten päähäni sytyttämä ajatuksen kipinä suurien ilmiöiden ymmärtämisestä on kenties ohjannut minua osaltaan maantieteen ja tähän työhön valitsemani aiheen pariin, vaikka valitsin työn aiheen täysin eri syistä kuin halusta ottaa pieni askel kohti Säätiö-kirjasarjan maailmaa. Kaupungin asukkaiden keskinäinen vuorovaikutus yksilöinä ja ryhminä, heidän välillä keskustelujen kautta liikkuvat ajatukset ja kokemukset elinympäristöstään sekä tapa miten fyysinen kaupunkitila voi näyttäytyä eri ihmisille täysin eri lailla ovat kiinnostaneet minua varsin pitkään. Tätä tekstiä kirjoittaessani Asimovin kirjasarja muistui jatkuvasti mieleeni ja aloitinkin kirjasarjan kevyen lukemisen vastapainona varsinaiselle tutkimus- ja kirjoitustyölle. Näistä syistä, Isaac Asimovin Säätiö-kirjasarjan mainitseminen lyhyesti tekstin alussa on mielestäni osuvaa, vaikka tekstin aihe ei käsittelekään mitään psykohistorian kaltaista asiaa.

Tämä opinnäytetyö on pitkän työn tulos, josta haluan kiittää lukuisia henkilöitä. Erityisesti haluan kiittää ohjaajiani geoinformatiikan apulaisprofessori Tuuli Toivosta sekä englannin kielen ja digitaalisten ihmistieteiden apulaisprofessori Tuomo Hiippalaa heidän ohjauksestaan, kommenteistaan ja panoksestaan tämän työn loppuun viemisessä. Lisäksi haluan kiittää Henrikki Tenkasta python-opetuksesta, Miikka Silfverbergiä FinnPOS-opastuksesta, Elias Willbergiä ideoiden pallottelusta ja vertaistuesta sekä kaikkia Digital Geography Labin tutkijoita, joilta olen pyytänyt mielipiteitä visualisointeihin liittyen. Erityisen suuri kiitos puolisolleni Mimosalle, joka on jaksanut tukea ja kuunnella minua tämän työn aikana, sekä kannustanut panostamaan työhön alusta loppuun sataprosenttisesti.

Sanasto

Aihemallinnus (engl. topic modeling) = Kieliteknologinen tietokoneavusteinen menetelmien kehikko, jonka yleisenä toimintaperiaatteena on louhia sille annetuista teksteistä niissä ilmeneviä aiheita. Aihemallinnusta voi tehdä usealla eri menetelmällä, mutta käytetyin algoritmi on LDA.

Emoji = Tekstissä esiintyvä pienikokoinen kuva, jota käytetään tehostamaan tekstin tunneviestiä. Hymiö eroaa emojista siten, että se ei ole kuva vaan toteutettu kirjaimien ja välimerkkkien yhdistelmistä.

Geoleimaus (engl. geotagging) = Maantieteellistä sijaintia kuvaavan tiedon lisääminen esimerkiksi valokuvaan, videoon tai äänitiedostoon. Geoleimaus mahdollistaa esimerkiksi sosiaalisessa mediassa tuotetun tiedon tarkastelun spatiaalisesti. Instagramissa geoleimaukset sidotaan tietokantaan tallennettuihin kohdepisteisiin.

Hukkasana (engl. stop word) = Tietosisällöltään köyhä ja analyysin kannalta hyödytön sana, joka jätetään analyysin ulkopuolelle tuloksien parantamiseksi. Näitä ovat muun muassa “a”, “the”, “että” sekä “ja”. (Koskenniemi 2011)

Kieliteknologia (engl. linguistic technology) = Koneoppimista ja tietojenkäsittelymenetelmiä yhdistelevä metodologia, jossa käsitellään ja analysoidaan ihmisten käyttämillä kielillä koostettuja aineistoja kuten tekstiä ja puhetta. Kieliteknologiaa käytetään yleisesti esimerkiksi aihemallinnukseen, sävyanalyysiin ja lemmatisointiin.

Kohdepiste (engl. point-of-interest) = Kohdepiste on johonkin maantieteelliseen sijaintiin sidottu piste, jonka kohde voi olla mielenkiintoinen tai hyödyllinen kontekstista riippuen. Internetin karttapalveluissa paikat ovat usein kohdepisteitä.

LDA (Latent Dirichlet Allocation) = LDA on yleisesti käytetty aihemallinnusmenetelmä, joka perustuu oletukseen siitä, että jokainen teksti koostuu erilaisten aiheiden sekoituksista ja tekstin jokaisen sanan voi luokitella jonkin aiheen alle. Aiheet muodostuvat samojen sanojen esiintymisistä lähekkäin tekstissä. Menetelmä on valvomaton, joten se tuottaa itse käyttäjän määrittelemän määrän aiheita.

Lemmatisointi (engl. lemmatization) = Lemmatisointi on luonnollisen kielen prosessi, jossa sanan taivutusmuoto muutetaan sanan perusmuotoon. Lemmatisoinnissa pyritään huomioimaan sanan sijainti lauserakenteessa ja konteksti, jolloin sanan lemmatisoimisen onnistuminen on todennäköisempää. Esimerkiksi “Minun kotona oli vieraita viikonloppuna” muuttuu “minä koti olla vieras viikonloppu”. Tässä työssä käytettiin FinnPOS- ja spaCy-työkaluja suomeksi ja englanniksi kirjoitettujen aineistojen lemmatisoinneissa.

Mainінta (engl. mention) = Useissa sosiaalisen median alustoissa käytössä oleva toiminto, jossa julkaisun tai kommentin voi liittää toiseen käyttäjään. Mainінta on usein muodossa @käyttäjänimi. Esimerkiksi “Hei @kayttaja1, oletko jo maistanut @yritysXYZ sitruunasorbettia?”

Massadata (engl. big data) = Massadata on nykyaikainen käsite, joka kuvaa internetissä käyttäjien ja laitteiden tuottamaa dataa. Massadatalle ominaista on datan suuret määrät, nopea syntymistahti ja monipuolinen luonne.

Stemmaus (engl. stemming) = Lemmatisoinnin kaltainen, mutta karumpi ja mekanistisempi menetelmä, jossa taipuneiden sanojen päätteet poistetaan. Esimerkiksi “kotini” → “koti” ja “kodissani” → “kodi”. Toisin kuin lemmatisoinnissa, stemmauksessa sanaa ei palauteta sen perusmuotoon vaan siitä poistetaan pääte.

Vektorirepresentaatio (engl. vector representation) = Tekstimuotoiselle sanalle kehitetty numeerinen esitysmuoto, joka mahdollistaa tekstiaineiston laskennallisen käsittelyn.

Kuvat ja taulukot

Kuva 1	*s13*	Kuva 28	*s79*	Kuva 55	*s129*
Kuva 2	*18*	Kuva 29	*79*	Kuva 56	*132*
Kuva 3	*33*	Kuva 30	*80*	Taulukko 1	*15*
Kuva 4	*35*	Kuva 31	*82*	Taulukko 2	*49*
Kuva 5	*37*	Kuva 32	*83*	Taulukko 3	*53*
Kuva 6	*38*	Kuva 33	*85*	Taulukko 4	*55*
Kuva 7	*39*	Kuva 34	*85*	Taulukko 5	*56*
Kuva 8	*43*	Kuva 35	*91*	Taulukko 6	*59*
Kuva 9	*43*	Kuva 36	*92*	Taulukko 7	*61*
Kuva 10	*45*	Kuva 37	*94*	Taulukko 8	*64*
Kuva 11	*46*	Kuva 38	*95*	Taulukko 9	*67*
Kuva 12	*48*	Kuva 39	*96*	Taulukko 10	*88*
Kuva 13	*51*	Kuva 40	*97*	Taulukko 11	*89*
Kuva 14	*52*	Kuva 41	*99*	Taulukko 12	*106*
Kuva 15	*53*	Kuva 42	*100*	Taulukko 13	*107*
Kuva 16	*65*	Kuva 43	*102*	Taulukko 14	*109*
Kuva 17	*69*	Kuva 44	*104*	Taulukko 15	*112*
Kuva 18	*70*	Kuva 45	*105*	Taulukko 16	*112*
Kuva 19	*71*	Kuva 46	*108*	Taulukko 17	*116*
Kuva 20	*72*	Kuva 47	*110*	Taulukko 18	*116*
Kuva 21	*73*	Kuva 48	*113*	Taulukko 19	*121*
Kuva 22	*74*	Kuva 49	*115*	Taulukko 20	*121*
Kuva 23	*75*	Kuva 50	*118*	Taulukko 21	*126*
Kuva 24	*76*	Kuva 51	*119*	Taulukko 22	*127*
Kuva 25	*77*	Kuva 52	*123*	Taulukko 23	*131*
Kuva 26	*78*	Kuva 53	*125*	Taulukko 24	131
Kuva 27	*79*	Kuva 54	*128*	Taulukko 25	135

1.0 Johdanto

Tietokoneiden, älylaitteiden, internet-yhteyksien ja sosiaalisen median alustojen myötä suuri osa yhteiskuntaa ja ihmisten välistä kanssakäymistä on saanut digitaalisen ulottuvuuden. Arkipäiväiset asiat hoidetaan yhä useammin verkossa. Tämä ihmisten digitaalisessa maailmassa alati yleistyvä toiminta tuottaa suuren ja voimakkaasti kasvavan määrän monipuolista tietoa ihmisistä, yhteiskunnista ja useasta muusta asiasta. Tämänkaltaista tietoa kutsutaan massadataksi (engl. *big data*), jota syntyy myös passiivisesti käyttäessämme erinäisiä älylaitteita, kuten älypuhelimia. Toisin kuin esimerkiksi kyselytieto, massadataa syntyy jatkuvasti ja tieto tulee suoraan käyttäjiltä itseltään eikä "välikäsien", kuten erillisten kyselyjen, kautta. Erityisesti ihmisten itse tuottama tieto sosiaalisen median alustoilla tarjoaa mielenkiintoisen mahdollisuuden tutkia ja ymmärtää nykyaikaista maailmaa heidän näkökulmastaan. Tämänkaltainen tieto on erityisen mielenkiintoista maantieteen kannalta, sillä siihen usein liittyy jokin fyysisessä maailmassa oleva paikka koordinaattitietojen ja paikannimien muodossa. Digitalisoituvan maailman ymmärtäminen on erityisen tärkeää, sillä useat yhteiskunnan toiminnot ja ihmisten välinen kanssakäyminen hyödyntävät digitaalisia teknologioita tai joissain tapauksissa tukeutuvat niihin miltei kokonaisuudessaan. Mikäli digitaalisessa maailmassa tapahtuvia ilmiöitä ja trendejä ei tutkita ja siten saavuteta uutta ymmärrystä muuttuneesta maailmasta, maantieteen, sekä tieteen ylipäätänsä, selitysvoima ja uskottavuus luotettavan tiedon ja ymmärryksen tuottajana kärsii. Internetin tietokantoihin tallentuva älylaitteiden ja käyttäjien itse tuottama massadata, esimerkiksi sosiaalisen median julkaisujen, blogitekstien, videoiden ja äänitiedostojen muodossa, on huomioitu laajalti maailmalla, mutta rajoitetusti Suomessa.

Massadatan luonne edellyttää uusia poikkitieteellisiä menetelmiä, jotka mahdollistavat jaajojen ja monipuolisten aineistojen analyysin. Esimerkiksi reaaliaikainen tiedonkeruu kaupunkisuunnittelun ja päätöksenteon tueksi ja ilmiöiden seuraamiseksi on massadatavirtojen avulla täysin mahdollista, sillä massadataa syntyy jatkuvasti suuria määriä ja datavirtoja pystyy koostamaan sekä visualisoimaan sen syntyessä. On varsin selvää, että esimerkiksi jo olemassa olevien miljoonien sosiaalisen median julkaisujen käsittely ja analysointi kuvineen, kuvateksteineen ja kommentteineen puhtaasti "manuaalisesti" ei ole varteenotettava vaihtoehto huomioimatta lainkaan sitä, että kyseistä aineistoa syntyy päivittäin lisää satojen

tuhansien tai miljoonien julkaisujen muodossa. Massadata vaatii massadatalle soveltuvia nopeita ja tehokkaita käsittely- ja analysointimenetelmiä, kuten tekoälyyn ja koneoppimiseen pohjautuvia menetelmiä. Nykyaikaiset tekoälyä ja koneoppimista hyödyntävät menetelmät mahdollistavat suurten tietoaineistojen läpikäymisen ja luokittelemisen huomattavan nopeasti. Tällöin päästään käsiksi numeeristen tietomäärien lisäksi myös tiedon sisältöön ja näiden menetelmien voisikin sanoa mahdollistavan kvantitatiivisten ja kvalitatiivisten menetelmien hybridisoitumisen ainakin tietyissä skenaarioissa (Martin & Schuurman 2017). Tässä työssä pyritään sosiaalisen median aineistoja ja koneoppimista hyödyntäviä kielitieteellisiä menetelmiä käyttäen saamaan kuva suomen- ja englanninkielisten Instagram-julkaisujen kautta välittyvästä Helsingin digitaalisesta kaupunkitilasta ja ymmärtämään minkälainen digitaalinen kaupunkitila Helsingissä oikeastaan on.

Tässä työssä hyödynnetään suositussa sosiaalisen median alustassa, Instagramissa, tuotettua massadataa Helsingin alueelta aineistona, josta mallinnettujen aiheiden avulla pyritään ymmärtämään Helsingin digitaalista kaupunkitilaa. Työssä keskitytään suomen- ja englanninkielisten Instagram-julkaisujen kuvateksteihin sekä niissä esiintyviin aihe-eroihin alueellisesti kielten välillä, mutta erittäin kevyesti myös kuvatekstien sävyihin. Julkaisuihin liittyvien kuvien, kommenttien ja muiden elementtien (esimerkiksi tykkäykset) analysointi jää tämän työn ulkopuolelle. Sosiaalisen median käyttäjien kielen valintaa julkaisuja tehdessä ja valinnan vaikutusta julkaisujen sisältöön ei ole vielä kirjoitushetkellä tutkittu kovin laajalti, eikä kielen valinnan taustatekijöitä käsitellä tässä työssä muuten kuin lähinnä spekulatiivisella tasolla. Nuorten suomalaisten osalta on tosin todettu, että he sekoittavat englantia ja suomea keskenään arkikeskusteluissa jatkuvasti ja tekevät suunnilleen yhtä paljon sosiaalisen median julkaisuja niin suomeksi kuin englanniksi (Leppänen et al. 2011, Hiippala et al. 2018). Perusoletuksena on, että suurin osa julkaisujen kuvateksteistä pitää sisällään kyseisen julkaisun aiheen, joka täten on mallinnettavissa. Lisäoletuksena on, että suomen- ja englanninkielisten Instagram-julkaisujen kuvateksteissä on ainakin hieman eroja aihesisällössä, erityisesti sen vuoksi, että englantia käyttävät suomalaisten lisäksi myös Helsingissä vierailevat matkailijat. Helsingin alueella asuvat ja Helsingissä vierailevat matkailijat keskittyvät kaupungissa liikkuessaan todennäköisesti osittain eri asioihin. Matkailija havainnoi enemmän ympäristöä ja muiden ihmisten sosiaalista

kanssakäymistä etsien niistä mielenkiintoisia piirteitä, kun taas paikallinen ei näe arkiympäristössään mitään erityistä vaan keskittyy arkielämäänsä liittyviin asioihin (Urry & Larsen 2011, 13-18). Näin ollen, on varsin todennäköistä, että Instagram-julkaisuista paljastuvissa aiheissa on vähintään jonkin verran eroja suomeksi ja englanniksi kirjoitettujen julkaisujen välillä, jotka kertovat eroista digitaalisen kaupunkitilan näkymisessä eri kieliryhmille.

Käytössä oleva aineisto kattaa kaikki geoleimatut Instagram-julkaisut Helsingistä vuoden 2015 alusta vuoden 2016 maaliskuun loppuun, joita on reilu 800 000 kappaletta. Erityiesti tarkastelussa ovat erityisesti suomen- ja englanninkielisten julkaisujen kuvatekstit, joita käsittelemällä ja mallintamalla pyritään vastaamaan seuraaviin kysymyksiin

1. Mistä aiheista Instagram-julkaisuja tehdään Helsingissä?
2. Onko Instagram-julkaisujen aiheissa alueellisia eroja?
3. Onko suomen- ja englanninkielisissä Instagram-julkaisujen aiheissa spatio-temporaalisia ja sisällöllisiä eroja?
4. Soveltuvatko kieliteknologiset menetelmät kaupunkitilan tutkimukseen?

Näihin kysymyksiin vastauksen saaminen vaatii laajamittaisen kuvatekstiaineiston esikäsittelyn. Jotta aineiston jakaminen suomen- ja englanninkielisiin julkaisuihin on mahdollista, käsittelemätömälle aineistolle täytyy suorittaa automaattinen kielentunnistus, sillä aineistossa ei itsessään ole käytettyjä kieliä indikoivaa ominaisuustietoa. Ennen kielentunnistustusta kuvatekstit siivotaan niissä esiintyvästä “hälinästä” kuten muun muassa maininnoista, hymiöistä, emojeista ja aihetunnisteista, lopulta jäljelle jää pelkästään itse kuvateksti. Automaattisen kielentunnistuksen jälkeen aineistosta erotellaan suomenkieliset ja englanninkieliset julkaisut kahdeksi, toisistaan erillisiksi, aineistoiksi. Näitä kahta aineistoa työstetään tämän jälkeen rinnakkain kullekin kielelle kehitetyillä työkaluilla. Molemmille aineistoille erottelun jälkeen lemmatisointi, eli sanojen muuttaminen niiden perusmuotoon, jolla pyritään parantamaan aihemallintamisen tuloksia. Aihemallintamisen jälkeen tulokset esitetään, analysoidaan ja niiden perusteella vastataan yllä esitettyihin tutkimuskysymyksiin. Tässä kappaleessa esitetyt toimenpiteet tehdään pitkälti Python-ohjelmointiympäristössä hyödyntäen Python-kielelle kirjoitettuja kirjastoja. Lopulta tulokset visualisoidaan paikkatieto- ja tilasto-ohjelmien avulla. Tulosten visualisoimisen ja esittämisen jälkeen niitä

hyödynnetään Instagramin kautta välittyvän digitaalisen kaupunkitilan hahmottamiseen ja erittelyyn, sekä julkaisukielien väliseen vertailuun.

Opinnäytetyön rakenne on seuraava: ensimmäinen kappale johdattelee aihepiiriin, taustoittaa työtä tutkimuskirjallisuuden avulla ja kytkee työn teoreettiseen keskusteluun digitaalisesta kaupunkitilasta. Toinen kappale esittelee aineiston ja menetelmät. Kolmannessa kappaleessa analysoidaan tuloksia taulukoin, kuvaajin ja kartoin. Neljännessä kappaleessa työn tulokset kytketään ensimmäisessä kappaleessa avattuun teoreettiseen keskusteluun ja pohditaan tuloksien merkityksiä niiden valossa.

1.1 Työn taustoittaminen

Ubiikki tietokoneistuminen, sosiaalisen median suosio ja alati halpenevat sekä nopeutuvat internet-yhteydet muokkaavat yhteiskuntaa ennen näkemättömin tavoin: älylaitteilla voi olla välittömässä yhteydessä lähes kaikkialle maailmaan ajasta ja lähes mistä tahansa sijainnista riippumatta. Internetissä sijaitsevat sosiaalisen median palvelut ovat mahdollistaneet laajamittaisen sosiaalisen verkostoitumisen, joka oli vielä hieman yli kymmenen vuotta sitten ennenkuulumatonta. Jatkuvasti internetiin yhteydessä olevat älylaitteet valtaavat ihmisten arkielämän puhelimista ja televisioista autoihin ja jääkaappeihin. Tämän valtauksen myötä ja näiden laitteiden ja niille kehitettyjen sovellusten avulla ihmisten arkielämä ja yhteiskunta muuttuvat: reaaliaikaiset reittiohjeet ja aikataulutiedot mahdollistavat tehokkaan liikkumisen arkiympäristössä, uudet ja jännittävät asiat sekä ilmiöt leviävät yleiseen (jopa globaaliin) tietoisuuteen nopeasti ja koko arkiympäristön voi valjastaa muun muassa peliareenaksi. Älylaitteet voivat keskustella toistensa kanssa jolloin esimerkiksi jääkaappi voi automaattisesti ilmoittaa omistajansa älypuhelimeen maidon olevan lopussa tai älykäs kukkaruukku ilmoittaa kastelujärjestelmälle tarvitsevansa lisää vettä.

Useat sosiaalisen median palvelut, kuten Facebook, Instagram ja Twitter, mahdollistavat vapaan itseilmaisun tekstien, kuvien, äänen ja videoiden myötä, joiden myötä ihmisten ajatukset ja kokemukset eivät välttämättä ole vain rajatun joukon tiedossa vaan yhtenä julkaisuna julkaisuvirrassa ja kenen tahansa tarkasteltavissa. Tämä on mullistavaa, sillä pitkään julkisen keskustelun portinvartijoina olivat perinteinen media, lehdistö ja tv-kanavat (Papacharissi 2002; Brants 2005). Sosiaalinen media on täynnä julkaisuja, jotka tulevat

periaatteessa suoraan käyttäjiltä eivätkä ole muiden sensuroimia käyttäjää tai alustaa lukuun ottamatta. Julkaisuihin usein liittyy spatiaalista tietoa eli esimerkiksi kuvan maantieteelliset koordinaatit sekä temporaalista tietoa kuten päivämäärä sekä kellonaika. Tämä tekee sosiaalisen median käyttäjien julkaisuista arvokkaan tiedonlähteen kulttuuria, yhteiskuntaa ja siinä eläviä ihmisiä tutkittaessa. Kaupungit ovat erittäin luontevia kohteita sosiaalisen median tutkimukselle, sillä ihmiset ovat keskittyneet niihin, jonka vuoksi myös kaupunkialueilla on luonnonalueita suurempi määrä sosiaalisen median julkaisuja. Sosiaalisesta mediasta louhitun tiedon hyödyntäminen yhteiskunnalliseen keskusteluun vaikuttamisessa on ollut vuoden 2018 aikana vahvasti uutisissa (Cadwalladr & Graham-Harrison 2018, Mäkinen 2018).

Kaupunkien “syke” on tärkeä ihmisiä kaupunkeihin houkutteleva tekijä (Townsend 2013, 1-4). Istuminen vilkkaan kadun varrella kahvilassa voi saada aikaan tunteen siitä, että pitäisi sormiaan ”kaupungin valtimolla” ja pystyisi aistimaan sen ilmapiirin ja asukkaiden ajatukset. Sormensa laittaminen kaupungin sydämensykkeen päälle on aivan viime vuosiin asti ollut lähinnä runollinen vertauskuva. Tätä on romantisoitu lukuisissa kirjoissa, sarjakuvissa, elokuvissa ja tv-sarjoissa, jonka myötä siitä on tullut eräänlainen “performanssi” tai symboli kaupungissa asumiselle ja urbaanille ihmiselle. Se on kuvattu eräänlaisena todellisen kaupunkilaisuuden piirteenä, jonka tavoittelu on kaupunkilaiselle itseisarvo. Ilmiö on kuin käänteinen versio “luonnon kanssa harmoniassa olemisesta”, jossa henkilö osaa lukea luonnonympäristöään ja kokee olevansa siihen “kytkettynä” (Townsend 2013, 1-4). Informaatioteknologian, sosiaalisen median ja ubiikin tietokoneistumisen myötä on tullut mahdolliseksi “laittaa sormi kaupungin sykkeelle” ainakin osittain (Batty 2010). Esimerkiksi seuraamalla ”urbaanien kojelautojen” (kuva 1) reaaliaikaisesti visualisoimia tietovirtoja voi käytännössä saada edellä kuvatun kaltaisen romantisoidun urbaanin kokemuksen, mutta henkilökohtaisten tuntemuksien sijasta kokemuksen avaimena toimii datavirroista, sensoreista ja mittareista kerätty tieto. Tämän työn aihe on ajankohtainen, sillä työn kirjoitushetkellä kansainvälisissä tiedotusvälineissä ja sosiaalisessa mediassa keskustellaan sosiaalisen median tietojen hyödyntämisessä vaalimainonnan kohdistamisessa tiettyihin ihmisryhmiin, jonka myötä erään suurimman sosiaalisen median alustan, Facebookin, omistaja Mark Zuckerberg on ollut Yhdysvaltain

senaatin kuulusteltavana alustan roolista ja vastuusta Yhdysvaltojen ulkopuolisten tahojen ajaman vaalivaikuttamisen kitkemisessä (Wylie 2018; Wong 2018).

Kuva 1. Rio De Janeiron operatiivinen keskus, ääriesimerkki urbaanista kojelaudasta, jossa reaaliaikaista tietoa kaupungista visualisoidaan, seurataan ja sen perusteella tehdään päätöksiä (Places 2015).

Instagram-aineistossa esiintyvien käyttäjien yksityisyyden suojelemiseksi kenenkään käyttäjän oikeaa nimeä, käyttäjänimeä ei julkaista tässä työssä. Esimerkkeinä käytetyt kuvatekstit eivät ole oikeita kuvatekstejä vaan oikeisiin kuvateksteihin perustuvia ja niitä mukailevia esimerkkitekstejä, jotka ovat anonymisoitu Markhamin (2012) ehdottamalla tavalla. Kartoissa käytetään pääsääntöisesti ruudukko- tai kennostotasoja tietojen visualisointiin, jolloin yksittäiset julkaisut ja niiden tarkat koordinaatit eivät välity.

1.2 Kaupunkitila

Tila on eräs maantieteen perinteisimpiä käsitteitä, jota on käsitteellistetty yhä uudelleen ja uudelleen erityisesti 1900-luvun loppupuoliskolta lähtien, jolloin käsitys tilasta absoluuttisena ja mitattavana “säiliönä”, jossa tapahtuu erilaisia aktiviteetteja, alkoi muuttua (Castree et al. 2013). Edellä mainitun absoluuttisen tilan rinnalle nousi kognitiivisesti hahmotetun tila käsite, jonka mukaan ihmiset asuvat ja elävät absoluuttisessa tilassa, mutta he eivät ajattele sitä absoluuttisena tilana, vaan omien tunteiden ja kokemustensa kautta. Tätä havainnollistaa perinteikäs esimerkki: sama silta tuntuu kävelijästä huomattavasti pidemmältä kylmällä

sadesäällä kuin tyynenä ja lämpöisenä kesäpäivänä. Tämän rinnalle nousi relationaalisen tilan käsite, joka on kriittisten maantieteilijöiden näkemys tilasta. Siinä tila ei ole koskaan neutraali ja passiivinen geometrinen asia, vaan sosiaalisten suhteiden ja käytäntöjen jatkuvasti tuottama ja muokkaama tila, joka muistuttaa huomattavasti julkisen tilan käsitettä (Gohen 1998), tosin sisältäen myös yksityiset tilat. Feministisessä maantieteessä tätä relationaalisen tilan käsitettä vietiin pidemmälle, jolloin tuotettu tila voi olla moni-merkityksellistä, paradoksaalista ja sekavaa. Castreen ja muiden (2013) määritelmän mukaan nykyaikaisin käsitys tilasta määrittelee tilan siten, että se muotoutuu jatkuvasti tilan käytöstä ja käytänteistä, eikä sitä voi täten määritellä kiinteällä määritelmällä. Kaupunkitilalla tässä yhteydessä tarkoitetaan pitkälti tilan käsitteen siirtämistä kaupunkiympäristöön, jossa se ymmärretään yllämainitun nykyaikaisimman määritelmän kautta kaupungissa olevana tilana, joka syntyy tilassa tapahtuvista toiminnoista ja käytänteistä.

Tanskalainen arkkitehti Jan Gehl on kirjoittanut kaupunkitilojen merkityksestä ja laadukkaasta kaupunkiympäristöstä 1970-luvulta asti. Kirjassaan *Life between buildings* hän jakaa kaupungissa sijaitsevien rakennusten välisen tilan eli katutilan kolmeen eri päätyyppiin niissä harjoitettavien aktiviteettien mukaisesti (Gehl 2011). Niitä ovat: välttämättömät aktiviteetit, valinnaiset aktiviteetit ja sosiaaliset aktiviteetit. Nämä aktiviteetit asettavat kukin fyysisen ympäristön rakenteelle tiettyjä vaatimuksia. Välttämättömät aktiviteetit ovat arkipäiväisiä aktiviteettejä kuten töissä ja ostoksilla käyminen, ylipäätään arkielämän kannalta välttämätön ”asioiminen”. Hän sanoo, ”ettei nämä välttämättömät aktiviteetit aseta ympäristölle juurikaan vaatimuksia, sillä niitä tullaan tekemään päivittäin läpi vuoden” (Gehl 2011). Valinnaiset aktiviteetit ovat nimensä mukaisesti sellaisia, joita henkilö voi valita tekevänsä tai jättää tekemättä ajan ja ympäristön niin salliessa. Esimerkkejä tällaisista aktiviteeteistä ovat muun muassa kävelyllä käyminen, kahvin nauttiminen terassilla ja auringonotto. Gehlin mukaan ihmiset eivät harrasta valinnaisia aktiviteettejä, mikäli ulkoiset tekijät, kuten sää ja paikka, eivät ole suotuisia tai henkilölle mieluisia (Gehl 2011). Tosin myös sisäisillä tekijöillä, kuten ihmisen mielialalla, kiistämättä on vaikutus siihen, ryhtyykö kyseinen henkilö esimerkiksi ottamaan aurinkoa, vaikka hän olisikin aurinkoisena päivänä rannalla. Edellä mainitut välttämättömät ja valinnaiset aktiviteetit käsittävät ne aktiviteetit, mitä asukkaat tekevät yksilöinä itsekseen. Sosiaalisilla aktiviteeteillä Gehl (2011) tarkoittaa kaikkea kanssakäymistä, joka on riippuvainen muiden ihmisten

läsnäolosta kuten keskustelut, lasten leikit ja vastaantulijoiden tervehtiminen. Gehl (2011) toteaa näiden sosiaalisten aktiviteettien syntyvän spontaanisti välttämättömistä ja valinnaisista aktiviteeteista ihmisten liikkuessa samassa ympäristössä, sillä ihmisten välinen kanssakäyminen houkuttelee muita ihmisiä. Tämä aktiviteetteihin perustuva kaupunkiympäristön laadun kehikko on esitelty taulukossa 1.

Huonon laadun kaupunkitilat eivät mahdollista valinnaisia aktiviteetteja lainkaan, joten niissä tapahtuu vain välttämättömiä aktiviteetteja, kun taas hyvän laadun katutila mahdollistaa myös valinnaiset aktiviteetit. Toisin sanoen, ihmiset eivät oleskele tilassa, jossa tapahtuu vain välttämättömiä aktiviteetteja, vaan pyrkivät siirtymään niistä pois esimerkiksi kotiin tai miellyttävämpään tilaan. Hyvän laadun katutilassa myös välttämättömien aktiviteettien suorittaminen voi pidentyä ajallisesti ympäristön miellyttävyyden vuoksi.

Taulukko 1. Jan Gehlin luoma kaupunkiympäristön laatukehikko (Gehl 2011).

	Huono ympäristö	**Hyvä ympäristö**
Välttämättömät aktiviteetit	Vähän	Vähän
Valinnaiset aktiviteetit	Vähän	Paljon
Sosiaaliset aktiviteetit	Vähän	Keskiverto

Jan Gehlin katutilan luokittelun punaisena viivana on ihmisten välisen kanssakäymisen mahdollistaminen, joka puolestaan luo hyvän katutilan ja kaupunkiympäristön. Hänen mukaansa hyvän kaupunkitilan merkki on se, kun suuri määrä ihmisiä viettää siellä aikaa muutenkin kuin välttämättömien aktiviteettien merkeissä. Tämän ajatuksen perusteella voisi olettaa suuren Instagram-julkaisujen määrän kertovan ihmisten viettävän alueella aikaa muissakin kuin välttämättömien aktiviteettien merkeissä, sillä sosiaalisen median julkaisun tekeminen vaatii jonkin verran vapaa-aikaa ja aiheen, josta haluaa kertoa ystävilleen ja seuraajilleen sosiaalisessa mediassa. Näin ollen, suuren aktiivisuuden alueet kenties vastaavat laadukasta kaupunkitilaa, mutta ei välttämättä kerro suuresta aiheiden kirjosta.

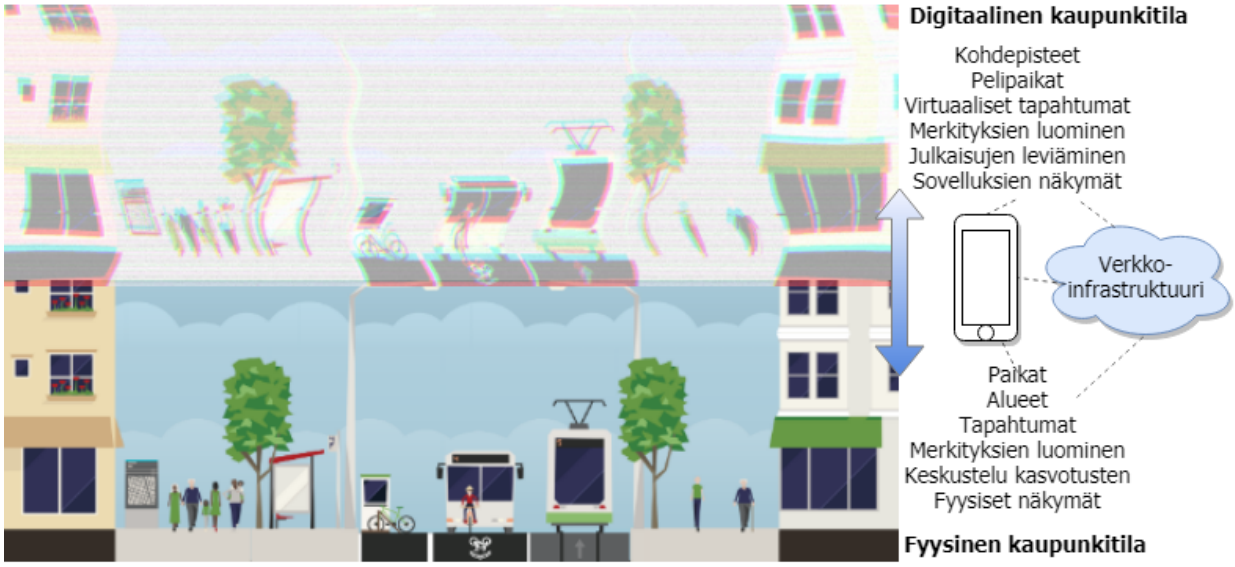
1.2.1 Digitaalinen kaupunkitila

Kaupunkitilan käsite laajenee huomioitaessa eletyn ja fyysisen kaupunkitilan lisäksi samoissa maantieteellisissä koordinaateissa sijaitseva paljaille silmille näkymätön kerros kaupunkitilassa, joka on saavutettavissa vain verkkoon kytkettyjen älylaitteiden ja internet-sovellusten kautta (kuva 2). Tämä uusi ulottuvuus kaupunkitilaan on syntynyt mobiililaitteiden, langattomien laajakaistayhteyksien ja sosiaalisen

median alustojen myötä (Zook & Graham 2007, 2017; Batty 2010; Graham & Zook 2011; Drucker & Gumpert 2012; Stefanidis et al. 2013; Kellerman 2014; Malecki 2017; Rose 2017). Digitaalinen kaupunkitila on näyttäytynyt kirjallisuudessa usealla eri nimellä ja muutamalla eri määritelmällä. Rose (2017) puhuu digitaalisesti välittyvästä kaupunkitilasta, Kellerman (2014) puhuu "tuplatilasta" ja Graham & Zook (2011) sekä Malecki (2017) puhuvat kyber- ja hybriditiloista sekä kybermaisemista, mutta he käsittelevät samaa asiaa: digitaalista kaupunkitilaa ja sen monimutkaista suhdetta fyysiseen kaupunkiin ja kaupunkitilaan. Digitaalista kaupunkitilaa melko lähellä oleva Stefanidiksen ja muiden (2013) kehittämä käsite "ympäröivä geospatiaalinen tieto" (engl. *ambient geospatial information*) keskittyy pitkälti sosiaalisen median aineistoihin lähes neutraaleina tiedostoina ja tietokantoina, joista voidaan johtaa fyysistä maailmaa kuvaavaa paikkatietoa. Sosiaalisen median aineistojen käsitteleminen neutraaleina, eikä käyttäjiensä ja käyttökontekstien kirjoa huomioivina monimutkaisina sekä ristiriitaisina aineistoina, on turhan yksinkertainen näkemys digitaalisesta kaupunkitilasta (Rose 2017). Kaupunkitila koetaan, eletään ja keksitään uudelleen digitaalisessa maailmassa alati kasvavissa määrin älylaitteiden ja asioiden internetin yleistyessä (Rose 2018). Fyysisen ja digitaalisen kaupunkitilan suhteita toisiinsa värittää pitkälti internetin käyttäjäkunnan monimuotoisuus, esimerkiksi Rose (2018) esittää älylaitteiden käyttäjien ja sovellusten käyttötapojen olevan erittäin moninaisia, joten sama fyysinen kaupunkitila voidaan siten myös esittää, tuottaa ja kokea digitaalisesti lukuisin eri tavoin. Tämä näkemys digitaalisesta välitetystä kaupunkitilasta sisältää vahvoja viitteitä kriittisten ja feminististen maantieteilijöiden näkemykseen tilasta sosiaalisesti tuotettuna ja ristiriitaisena alati muutoksen alla olevana tilana. Näkemys tuntuu päällepäin varsin osuvalta ottaen huomioon kuinka moninainen joukko internetin ja sosiaalisen median käyttäjien joukko on, jolloin saman asian ristiriitaiset tulkinnat käyttäjäjoukon sisällä ovat lähes varmoja.

On epäselvää kuinka vahvasti ja millä lailla digitaalinen kaupunkitila ja fyysinen kaupunkitila eroavat toisistaan ja missä niiden välinen raja menee, mutta ne ovat kytkeytyneet toisiinsa (Kitchin 1998; Cohen 2007; Zook & Graham 2007, 2017; Kellerman 2011, 2014; Tranos & Nijkamp 2013; Malecki 2017; Rose 2017). Rajat digitaalisen ja fyysisen kaupunkitilan välillä ovat häivettyneet paikoin jo niin, että selkeätä rajaa on vaikea vetää (Malecki 2017). Esimerkiksi digitaalinen kaupunkitila on olemassa ja sinne pääsy on

mahdollista vain fyysisesti jossain sijaitsevan internet-verkkoinfrastruktuurin avulla: palvelimet, reitittimet, kaapelit ja langattomien yhteyksien asemat sijaitsevat jossain fyysisessä paikassa, eivätkä juurikaan muuta sijaintiaan (Tranos & Nijkamp 2013; Kellerman 2014; Malecki 2017). Lisäksi, digitaaliseen kaupunkitilaan pääsee käsiksi vain älylaitteen tai tietokoneen kautta, joka on yhdistetty verkkoon edellä mainitun infrastruktuurin kautta, mutta itse verkkoa tai digitaalista kaupunkitilaa ei voi havaita paljain silmin. Verkkoon yhdistäminen voi tietyissä fyysisen kaupunkitilan sijainneissa olla vaivalloista tai jopa mahdotonta, jos mobiiliverkossa on sijainnin kohdalla katvealue eikä sijainnissa sijaitsevat wifi-reitittimet tai verkkoon pääsy ylipäätään ole avoimia. Tällöin sijainnista itsestään ei pääse verkkoon ilman käyttäjätunnuksia ja sitä kautta sijainnin digitaalista kaupunkitilaa pääsee tarkastelemaan vain oltaessa fyysisesti sen ulkopuolella. Tilanne on verrattavissa julkisiin ja yksityisiin tiloihin fyysisessä kaupunkitilassa ja voi altistaa esimerkiksi syrjityt väestöryhmät digitaaliselle marginalisoitumiselle (Graham & Zook 2011, 2013; Tranos & Nijkmap 2013; Malecki 2017; Rose 2017), jolloin he eivät kaupunkitilan lisäksi näy myöskään digitaalisessa kaupunkitilassa. Internet-verkko itsessään ei sijaitse varsinaisesti missään, mutta samaan aikaan se sijaitsee kaikkialla ja on yhteydessä kaikkialle muualle ja katukuvassa näkyvien älylaitteiden kautta jokainen paikka on kytköksissä muihin paikkoihin lähes jatkuvasti (Rose 2017). Tämä johtaa tilanteeseen, jossa älylaitteen käyttäjä on yhteydessä verkkoon ja liikkuu siellä tarkastellen esimerkiksi lähialueelta löytyviä ravintoloita, samanaikaisesti hän liikkuu fyysisesti olemassa olevassa verkkoinfrastruktuurissa, verkossa eli kaikkialla ja ei missään, verkkosovelluksessa sekä digitaalisesti välitetyssä kaupunkitilassa samanaikaisesti (Kellerman 2014; Malecki 2017; Rose 2018). Digitaalinen kaupunkitila on siis varsin ristiriitainen käsite ja sen määritelmä hakee vielä muotoaan.



Kuva 2. Havainnollistus digitaalisen kaupunkitilan käsitteestä. Digitaalinen kaupunkitila heijastaa fyysistä kaupunkitilaa, on nähtävissä vain verkkoon kytkettyjen laitteiden kautta ja sijaitsee paradoksaalisesti samanaikaisesti useassa sijainnissa ja ei missään erityisessä sijainnissa. Katukuva on luotu Streetmix- ja Photomosh-internetsovelluksilla.

Internetin ja sosiaalisen median vaikutus elettyyn fyysiseen kaupunkiympäristöön voi olla suoraa, epäsuoraa ja vastavuoroista (Kellerman 2014; Rose 2017). Digitaalinen kaupunkitila vaikuttaa myös Jan Gehlin aktiviteettiperusteiseen kaupunkitilan jaotteluun, joka perustui pitkälti fyysisestä ympäristöstä kumpuavaan ihmisen aktiviteettivalintoihin vaikuttavaan ajatukseen. Digitaalisessa kaupunkitilassa ihmiset jakavat kokemuksiaan, luovat ja uudelleenluovat mitä erilaisempia merkityksiä täydentämään, muuttamaan tai korvaamaan niin digitaalista kuin fyysistä kaupunkitilaa (Kellerman 2014; Malecki 2017; Rose 2017). Toisin kuin fyysinen kaupunkitila, digitaalinen kaupunkitila voi muuttua nopeaan tahtiin esimerkiksi uusien internet-ilmiöiden myötä tai käyttäjien mielenkiinnonkohteiden muuttuessa: jokin paikka kaupungissa saattaa olla erittäin hiljainen, välttämättömien aktiviteettien kaupunkitila tavallisesti, mutta esimerkiksi sosiaalisessa mediassa tapahtuva tempaus voi saada ihmisvyöryn aikaan, jolloin tilan “aktiviteettiluonne” muuttuu hetkellisesti selkeästi. Esimerkiksi vuoden 2016 kesällä yleisölle avattu Vallisaari on todennäköisesti ollut lähes näkymätön sosiaalisen median alustoilla ennen avaamistaan ja avaamisen jälkeen Vallisaaren näkyvyys nousee erityisesti retkikohteena.

Riippuen sovelluksesta digitaalinen kaupunkitila luodaan joko ruohonjuuritasolta ylöspäin tai ylhäältä alaspäin. Esimerkiksi älylaitteiden ja -sovellusten hyödyntämän reaaliaikaisen sijaintitiedon ja käyttäjän intresseistä luodun profiilin perusteella digitaalisesta kaupunkitilasta voidaan “nostaa” tai “peittää” tiettyjä

piirteitä (Graham & Zook 2011; Dotson 2012; Malecki 2017; Rose 2017). Tällöin tarkastellessaan ympäristöään älylaitteen kautta, käyttäjä ei välttämättä voi vaikuttaa siihen, mitä älylaitteen ruudulla korostetaan tai jätetään kokonaan näkymättömiin, jolloin digitaalinen kaupunkitila muodostuu ylhäältä alaspäin. Sama pätee myös “neutraaleihin” karttapalveluihin, kuten Googlen kartta- ja navigointipalveluun, joka on erittäin laajalle levinnyt ja yleisesti käytössä oleva navigointisovellus, jossa päätökset siitä mitä kartalla näytetään tai ei näytetä on puhtaasti yhtiön käsissä (Malecki 2017). Esimerkiksi Googlen karttapalvelu tarjoaa osittain eri ravintoloita, kun käyttäjä tekee haun “ravintola helsinki” -haulla ja “restaurant helsinki” -haulla. Täysin päinvastainen esimerkki on OpenStreetMap, jossa käyttäjät itse toimivat sisällöntuottajina ja -valvojina, mutta tässäkin tapauksessa mitä kartoitetaan, mitä ei ja kuinka tarkasti on käyttäjä-sisällöntuottajista kiinni.

Käyttäjät voivat itse tuottaa digitaalisesti omia paikkojaan useiden älylaitesovelluksien muistiin ja myös jakaa näitä paikkoja muille käyttäjille (Rose 2017; Hochmair et al. 2018), jolloin fyysisen paikan digitaalisessa versiossa voi tapahtua merkityksen muutoksia (Rose 2017). Digitaalisessa kaupunkitilassa käyttäjät pystyvät muuttamaan fyysisen kaupunkitilan merkityksiä tekemällä niistä muun muassa peliareenoja, kohtaamispaikkoja tai statuksen nostamispaikkoja, vaikka itse sovellukset eivätkä fyysisen kaupunkitilan kohteet olisi suunniteltu näitä tarkoituksia varten lainkaan (Jenkins et al. 2016). Paikkojen merkityksen jatkuva muuttuminen ja muuttaminen voi olla erittäin voimaannuttavaa ja vapauttavaa henkilöille ja ryhmille, jotka ovat kokeneet vastoinkäymisiä ja voimattomuuden tunteita perinteisiä vaikutuskanavia pitkin (Kellerman 2014; Rose 2017). Digitaalisessa kaupunkitilassa liikkuminen ja toimiminen ovat pitkälti riippuvaisia internet-yhteyksien kattavuudesta ja toimivuudesta jolloin joillekin alueille ei mennä, jos siellä ei ole mobiilidata-yhteyksiä tai wifi-verkkoa tai ne ovat huonot (Rose 2017) ja käyttökatkosten sattuessa koko digitaalinen kaupunkitila “sammuu” tai “häviää”, joka voi näkyä fyysisessä kaupunkitilassa hetkellisesti hämmentyneinä henkilöinä. Tämä voi myös entisestään lisätä joidenkin internet-yhteyksien ulkopuolelle jäävien ihmisryhmien marginalisoitumista, jolloin vallan ja vaikutusmahdollisuuksien puutokset voivat näkyä entistäkin selkeämmin niitä tarkasteltaessa (Graham & Zook 2011; Tranos & Nijkamp 2013; Malecki 2017; Rose 2017). Käytetyn kielen vaikutusta marginalisoitumiseen digitaalisessa kaupunkitilassa on jonkin verran

näyttöä alueilta, joilla on eri väestöryhmien välisiä jännitteitä (Graham & Zook 2013). Tämän vuoksi kieliryhmien välinen tarkastelu on erityisen mielenkiintoinen tapa lähestyä digitaalista kaupunkitilaa myös Helsingissä.

Hieman käytäntöä lähemmällä tasolla esille nousee myös laitteiden näytöllä näkyvän tiedon priorisointi. Älylaitteiden näytölle ja sovellusten käyttäjänäkymään mahtuu kerrallaan vain rajallinen määrä tietoa ja se, mitä tietoa näytöllä näytetään, voi vaikuttaa vahvasti siihen, miten erilaiset alueet näyttäytyvät älylaitteiden käyttäjille (Malecki 2017). Sovelluksesta ja myös jopa käyttäjäprofiilista riippuen keskeisin tieto voi olla suoraan näytöltä nähtävissä, mutta jokin muu tieto voi olla useamman valikon ja "klikkailun" päässä, jota Kellerman (2015) kutsuu "digitaalisesti kauempana olemiseksi", vaikka fyysinen sijainti fyysisessä kaupunkitilassa ei muutu lainkaan. Kellerman (2015) myös väittää useiden sosiaalisten älypuhelinsovellusten käyttöliittymien olevan "epäpaikkoja" (engl. *non-place*) kuten fyysisen maailman "epäpaikat" eli paikat, jotka ovat kaikkialla samanlaisia huolimatta siitä missä maantieteellisessä sijainnissa ne sijaitsevat ja minä aikana niitä tarkastellaan (Castree et al. 2013). Näitä ovat perinteisen määritelmän mukaan muun muassa ostoskeskukset ja lentokoneasemat. Näkemyksessä on perää, sillä esimerkiksi Instagramin tai Twitterin käyttöliittymä on identtinen kaikilla käyttäjillä huolimatta siitä, onko käyttäjä Brasiliassa, Australiassa vai Suomessa, eli kyseessä on eräänlainen epätila. Tästä piirteestä huolimatta näissä käyttöliittymissä tehdyt julkaisut usein liittyvät fyysiseen maailmaan ja tiettyihin paikkoihin ja aikoihin fyysisessä maailmassa. Näin ollen kaupungissa liikkuvat älylaitteiden sovellusten käyttäjät ovat hieman paradoksaalisesti samanaikaisesti oikeassa fyysisessä tilassa, epäpaikassa ja digitaalisessa kaupunkitilassa (Kellerman 2015; Malecki 2017; Rose 2017).

Digitaalinen kaupunkitila näyttäytyy tieteellisen kirjallisuuden (Graham & Zook 2011; Kellerman 2014, 2015; Malecki 2017; Rose 2017) valossa varsin ristiriitaiselta, mutta kuitenkin ymmärrettävissä olevalta ja yllättävän selkeältäkin käsitteeltä. Sen moninaisuus ja ristiriitaisuus juontavat juurensa internetin käyttäjien moninaisuudesta ja sovellusten erilaisuudesta. Ilman Internet-verkkoa digitaalista kaupunkitilaa ei voi olla sellaisena kuin se on tässä työssä esitetty olevan, vaan molemmat ovat kytköksissä toisiinsa. Digitaalisen kaupunkitilan tarkkaa sijaintia on vaikea määritellä, koska verkon sijaintia on vaikea määritellä. Sen sijainti

on fyysisessä tilassa, epäpaikassa, verkko infrastruktuurissa ja digitaalisessa kaupunkitilassa itsessään, mutta samanaikaisesti se ei sijaitse missään näistä kokonaisuudessaan. Hahmottamisen kannalta digitaalisen kaupunkitilan voi ajatella olevan paljaille silmille näkymätön “kerrostuma” kaupunkitilassa, jossa tapahtuu aktiviteetteja, jotka osittain tukevat ja osittain korvaavat fyysisessä kaupunkitilassa tapahtuvia aktiviteetteja. Se on samanaikaisesti hierarkkisesti ja hajautetusti rakentunutta riippuen pitkälti sovelluksesta, jonka kautta digitaalista kaupunkitilaa tarkastellaan. Joka tapauksessa sosiaalinen media on mahdollistanut kaupunkitilan merkityksien muuttamisen ja demokratisoitumisen digitaalisessa kaupunkitilassa, josta muuttuneet merkitykset voivat valua ja näkyä myös fyysisessä kaupunkitilassa. Tämän myötä samalla paikalla voi olla lukuisia merkityksiä eri ihmisille ja merkitykset voivat olla toistensa kanssa ristiriidassa. Merkityksien moninaisuuden ja keskinäisen ristiriitaisuuden vuoksi kokonaisvaltaista kuvaa digitaalisesta kaupunkitilasta ei voi saada vain yhden älysovelluksen linssin läpi, mutta sen avulla voi saada kuvan kyseisen sovelluksen käyttäjien luomasta digitaalisesta kaupunkitilasta ja eri merkityksistä siellä.

1.2.2 Erilaiset ryhmät kaupungissa

Kaupungeissa asuu, liikkuu ja asioi suuri määrä ihmisiä, jotka muodostavat erilaisia ja eri kokoisia ryhmiä. Eräs tapa luokitella kaupungissa liikkuvat ihmiset ryhmiin yksinkertaisesti on jaotella ne paikallisiin ja vierailijoihin. Paikalliset ovat ihmisiä, jotka asuvat kaupungissa tai sen läheisyydessä ja joiden arjesta suuri osa kuluu kaupungin sisällä liikkumisessa ja asioimisessa. Vierailijat ovat ihmisiä, jotka eivät tavallisesti asioi kaupungissa osana arkeaan, jolloin siellä käyminen on heidän normaalista arjesta poikkeava tapahtuma, joka tehdään esimerkiksi vapaa-ajan tai pakollisen tarpeen vuoksi. Tapa, millä kaupunki näyttäytyy paikalliselle ja vierailijalle eroaa siten selkeästi toisistaan, sillä vierailijalle kaupunki on eksoottinen kohde ja paikalliselle osa arkiympäristöään, jonka hän ottaa niin sanotusti “annettuna”. Kaupungissa vierailulla olevat ihmiset eivät luultavasti eroa paikallisista järin suuresti siinä, minkälaisessa ympäristössä he tekevät mitäkin Gehlin aktiviteettilajia, mutta heidän aktiviteettilajit ovat todennäköisesti vahvasti kallellaan valinnaisiin ja sosiaalisiin aktiviteetteihin (taulukko 1). Kun vieraillaan toisessa kaupungissa, se tapahtuu usein vapaa-ajalla ja omasta tahdosta, jolloin ei luultavasti ole kiire arkiaskareiden pariin, kuten hakemaan lasta päivähoidosta tai viemään koiraa lenkille. Lisäksi matkailija on vieraassa kaupungissa ja kiinnittänee erityisesti huomiota siihen miltä eri ympäristöt näyttävät, kuulostavat ja tuoksuvat nauttiakseen vierailustaan ja vapaa-ajastaan,

sillä vierailu on selkeästi irtiotto tavallisesta arjesta eikä hän kulje ympäristössä normaaliin tapaansa (Urry & Larsen 2011: 3-10). Vierailija nimenomaan haluaa käydä paikoissa, jotka poikkeavat hänen tavanomaisesta ympäristöstä ja tutustua esimerkiksi erilaisiin aikakausiin historiallisesti tärkeiden alueiden ja paikkojen kautta. Matkailu antaa luvan ”kohdata ja vuorovaikuttaa ärsykkeiden kanssa, jotka luovat vahvan kontrastin arjen kanssa” (Urry & Larsen 2011: 3). Urryn ja Larsenin mukaan vierailija tavallaan kuluttaa matkakohteensa “ulkokuorta” (esimerkiksi arkkitehtuuria, ruokalajeja ja muiden ihmisten kanssakäymistä) aisteillaan. Hänen aistinsa aktiivisesti etsivät, arvioivat ja ottavat vastaan ympärillä tapahtuvia asioita elämyksinä arkisten tapahtumien sijaan.

Nykyaikaisessa matkailussa valokuvaus on erittäin yleistä, sillä se mahdollistaa vierailulla koettujen elämyksien kokemisen uudelleen ja jakamisen henkilöille ketkä eivät ole matkalla mukana, jonka myötä vierailijalla on syy ottaa lukuisia kuvia ympäristöstään. Sosiaalisen median alustat ovat erittäin suosittuja valokuvien jakamisen mahdollistavia palveluita, joihin lisätyt matkailukuvat toimivat matkailutoimiston kohde-esitteiden tavoin ja muokkaavat matkailijoiden olettamuksia ja toiveita kohteesta (Urry & Larsen 2011: 172-176, 186-188). Sosiaalisen median matkailukuvat voivat itseasiassa olla erittäin lähellä oikeita mainoskuvia, sillä on lähes itsestään selvää, että käyttäjät miettivät tarkkaan julkaisunsa visuaalisia piirteitä, jotta he pystyvät välittämään haluamansa mielikuvan omasta elämästään seuraajilleen (). Lisäksi usealla sosiaalisen median alustalla on “retrospektiiviominaisuus”, jossa käyttäjä esimerkiksi saa muistutuksen siitä, mitä teki vuosi sitten samana päivämääränä ja voi näin kokea matkan uudelleen. Lisäksi vierailijat todennäköisesti liikkuvat kaupungissa nähtävyyksien ja isoimpien kauppakatujen ja -keskuksien alueilla ja siten omassa ”kuplassaan”, jolloin heidän kokemus kaupungista muodostunee suurelta osin näiden alueiden pohjalta, josta on jonkin verran sosiaaliseen mediaan perustuvaa empiiristä näyttöä (Leung et al. 2017; Vu et al. 2017; Yang et al. 2017; Maeda et al. 2018). Tosin nykyisin osat vierailijoista ovat “post-turisteja”, jotka nimenomaan pyrkivät välttämään suurimpia turistinähtävyyksiä ja massatuotettuja turistielämyksiä, joiden sijaan he pyrkivät etsimään kohteestaan “autenttisia” paikkoja ja kokemuksia (Urry & Larsen 2011: 106-118).

Autenttisten matkakohteiden etsimisen suosion kasvu tapahtui samaan aikaan matkailun demokratisoitumisen kanssa. Autenttisuuden etsimisen myötä massaturismin suosituimpien kohteiden

vähätteleminen ja halveksunta sai alkunsa, jolloin “työväenluokan” matkakohteet nähtiin mauttomina massaturismin pilaamina aluein ja eksoottiset kohteet merkkinä matkailijan sosiaalisesta statuksesta ja menestyksestä (Urry & Larsen 2011: 32, 224-225). Tämänkaltaista kohteiden, tai paikkojen, välistä sosiaalisten hierarkioiden muodostumista tapahtuu myös nykyaikana varsin paljon. Hierarkioiden muodostuminen on todennäköisesti nopeutunut huomattavasti median demokratisoitumisen eli sosiaalisen median myötä, sillä kyseisissä medioissa yksityisihmiset ”esittelevät” omaa elämäänsä kuvin, videoin ja tekstein. Kun näitä julkaisuja kasautuu ajan saatossa suuria määriä, niin tietyt maat, kaupungit ja paikat korostuvat niissä eri lailla. Paikat, jotka saadaan näyttämään autenttisilta ja houkuttelevilta todennäköisesti nousevat hierarkiassa ylöspäin, kun taas tylsiltä ja epämieluisilta vaikuttavat paikat valuvat hierarkiassa alaspäin. Paikkojen hierarkkinen asema tosin riippuu monesta muustakin asiasta, kuin sosiaalisessa mediassa rakentuneesta mielikuvasta ja maineesta, esimerkiksi paikan täytyy myös vastata mielikuvia jossain määrin.

Paikallisen ja vierailijan välinen ero voi toimia eräänlaisena vertauskuvana eri kieliä käyttäville ryhmille sosiaalisissa medioissa ja digitaalisessa kaupunkitilassa. Muilla kielillä kuin suomeksi ja ruotsiksi tehdyt Instagram-julkaisut ovat melko todennäköisesti lähempänä vierailijaa, kuin paikallista. Tosin tämä kahtiajako on varsin karkea, eikä ota huomioon monikulttuurisia ja maahanmuuttotaustaisia paikallisia, saati sitä, että äidinkielenään suomea puhuva Instagram-käyttäjä voi tehdä julkaisuja, joiden kuvatekstit ovat jollain muulla kielellä kuin omalla äidinkielellään. Tästä piirteestä huolimatta on todennäköisempää, että puhtaasti suomeksi kirjoitetut Instagram-julkaisut ovat selkeästi vahvemmin paikallisten kuin matkailijoiden tekemiä verrattuna englanninkielisiin julkaisuihin. Englanti on erittäin yleinen kieli erityisesti Internetissä ja sen sovelluksissa, joten auttamatta myös suurehko osa paikallisia kirjoittaa julkaisunsa kuvatekstit englanniksi. Sama pätee myös vierailijoihin, joiden äidinkieli ei ole englanti, mutta he silti tekevät julkaisunsa englanniksi, jonka lisäksi englantia käyttävät myös sitä äidinkielenään puhuvat vierailijat. Englanninkielisiin julkaisuihin kuuluu siis todennäköisemmin suurempi osa vierailijoiden julkaisuja kuin suomenkielisiin julkaisuihin. Näin ollen eroja julkaisujen määrissä, sijainneissa ja aiheissa sekä aiheiden sijainneissa on oletettavissa löytyvän Instagram-aineistoa analysoitaessa suomenkielisten ja englanninkielisten aineistojen kautta. Tosin

paikallisiin tässä työssä luetaan myös suomeksi Instagram-julkaisuja tekevät, jotka asuvat muualla kuin pääkaupunkiseudulla, joten on mahdollista, että myös suomenkielisessä aineistossa paljastuu vierailijoiden intresseihin verrannollisia aiheita. Lisäksi näennäisesti aktiiviselta vaikuttava paikka tai alue Instagramissa ei välttämättä kerro myöskään sitä, että kyseessä olisi jonkin yhdistävän piirteen omaavan ihmisjoukon muodostavan yhteisön olemassaolosta ja liikkumisesta (Hochman & Manovich 2013).

1.3 Sosiaalinen massadata

Yhteiskunnassa tietoa tuotettiin, kerättiin ja lopulta tallennettiin pitkään lähinnä eri instituutioiden ja viranomaisten toimesta selvityksien, väestönlaskentojen tai muiden vastaavanlaisten hankkeiden yhteydessä. Hankkeiden välillä saattoi kulua parhaimmillaan useita vuosia tai vuosikymmeniä ennen seuraavaa kierrosta, jonka tuloksilla tietokantaa päivitettiin vastaamaan sen hetkistä tilannetta (Silva et al. 2014; Miller & Goodchild 2015). Suurten tietomäärien tuottamiseen ja keräämiseen ei nykypäivänä tarvitse valjastaa valtiollisia tahoja tai tutkimusinstituutteja tekemään pitkäkestoisia selvityksiä ja tiedonkeruun toimenpiteitä, sillä Web 2.0:n sovellukset kuten sosiaalisen median alustat ja asioiden internet (engl. *Internet of Things, IoT*) tuottavat automaattisesti suuria määriä tietoa. Esimerkiksi Google-hakuhistorioiden massadatasta louhittuja flunssa-aiheisia hakuja hyödyntämällä on ennakoitu lähes reaaliaikaisesti kausittaisen flunssan leviämistä (ns. “nowcasting”) ja sen on todettu olevan tarkkaa ja lähes reaaliaikaista (Graham & Zook 2011). “Perinteisin” keinoin kerättyä ja tuotettua tietoa on kutsuttu muun muassa “harvaksi dataksi” (Miller & Goodchild 2015). Harvan datan keräystä, tuotantoa ja tallennusta ei ole poistumassa, mutta sen rinnalle on noussut viranomaisten ja, ainakin näennäisesti, kolmansien osapuolien kontrollista vapaampi tiedon “kanava” internetin, laajakaistayhteyksien ja internetissä sijaitsevien sosiaalisten palvelujen yleistymisen myötä (Batty 2010; Malecki 2017).

Massadata on termi, jolla pyritään kuvaamaan nykyaikaisen yhteiskunnan, teknologian, kulttuurin ja näiden sisällä elävien yksilöiden tuottamaa jättimäistä tietovirtaa (Bender et al. 2014; Miller & Goodchild 2015; Rose 2017). Suurin osa etenkin länsimaisissa kaupungeissa asuvista ihmisistä ovat internetin käyttäjiä ja suuri osa internet-käyttäjistä ovat sosiaalisen median palvelujen, kuten Facebookin, Twitterin ja Instagramin, käyttäjiä. Näillä palveluilla on satoja miljoonia, ellei jopa yli miljardi käyttäjää alustaa kohden (Statista

2018a), joista kukin on sisällöntuottaja oman sosiaalisen median profiilinsa kautta. Sosiaalisen median palvelujen käyttäjät tekevät julkaisuja sosiaaliseen mediaan omasta elämästään, ajatuksistaan, mielenkiinnon kohteistaan ja unelmistaan toisten käyttäjien nähtäväksi ja kommentoitavaksi. Tämän sosiaalisesti tuotetun tiedon määrä on valtava, alati kasvava ja se tarjoaa mahdollisuuden päästä käsiksi sellaiseen tietoon nopeasti ja reaaliajassa, jota varten on aikaisemmin täytynyt käyttää hitaita ja pitkäkestoisia "harvaa dataa" keräämiä kysely- ja haastattelututkimuksia (Bender et al. 2014; Silva et al. 2014; Miller & Goodchild 2015). Esimerkiksi Foursquare- tai Swarm-alustojen paikkatietoa mielenkiintoisista paikoista, kohdepisteistä (engl. *point of interest, POI*), sekä käyttäjien kirjautumisista (engl. *check-in*) näihin paikkoihin on pystytty erittelemään kaupunkilaisten ruokailun ja vapaa-ajan tottumuksia (Silva et al. 2014). Sosiaalisen median julkaisutahdeissa näkyvät myös käyttäjien arkielämän rytmi ja erinäiset juhlapäivät aiheiden muodossa, mutta myös muutoksina käyttäjien julkaisutahdissa (Hochman & Manovich 2013). Eri sosiaalisen median alustoilla on toisistaan eriävät käyttäjäryhmät (Greenwood et al. 2016), jonka myötä useita eri sosiaalisen median palvelujen aineistoja yhdistelemällä pystytään mahdollisesti toteuttamaan entistä kattavampia analyysejä esimerkiksi tarkastelun alla olevan alueen, siellä liikkuvien ja asuvien ihmisten ominaisuuksista ja taipumuksista.

Massadata määritellään perinteisesti "neljän V:n" kautta (volume, variety, velocity, veracity) eli suomeksi vapaasti käännettynä määrän, monipuolisuuden, nopeuden ja totuudenmukaisuuden kautta (Bendler et al. 2014, Miller & Goodchild 2015). Sosiaalisen median ja asioiden internetin myötä tietoa on kertynyt ja kertyy suuria määriä, joista vasta murto-osaa on päästy analysoimaan tarkemmin. Yhä useamman henkilön liittyessä sosiaalisen median palveluihin sosiaalisen massadatan tuottajien määrä kasvaa. Tiedon laatu ja aiheet tosin ovat rajattuja sosiaalisen median alustan ominaisuuksiin ja rajoituksiin. Suuren datamäärän lisäksi massadataa määrittää sen datan tuottamisen nopeus. Tietoa syntyy nykymaailmassa jättimäisiä määriä päivittäin, esimerkiksi jo vuonna 2014 sitä arvioitiin syntyvän 2 biljoonaa gigatavua päivässä (Bender et al. 2014). Tämän määrä on oletettavasti kasvanut entisestään, sillä älylaitteet ja sosiaalisen median suosio ovat jatkaneet kasvuaan (Statista 2018a, 2018b, 2018c, 2018d). Älypuhelinten ja sosiaalisen median käyttäjien määrän kasvaessa nostaa uuden tiedon tuottamisen määrää jo itsessään, mutta sen lisäksi

jatkuvasti yleistyvät teräväpiirtovideot, 360 asteen panoraamakuvat ja lisätyn todellisuuden (engl. *augmented reality*) lisäävät tiedon määrää entisestään. Nämä edellä mainitut erilaiset mediat ovat muodostavat yhden massadatan määritelmistä, monipuolisuuden. Sosiaalinen massadata voi olla tekstiä, kuvia, videoita sekä näiden yhdistelmiä. On todennäköistä, että tietoa syntyy nopeammalla tahdilla kuin sitä ehditään eritellä, analysoida ja siitä ehditään tehdä johtopäätöksiä. Toisaalta sosiaalisen median julkaisut ovat niin monimuotoisia aiheiltaan, sisällöltään kuin media-alustaltaan, että niiden analysoiminen vaatii tarkan ja perusteellisen esikäsittelyn. Tähän esikäsittelyyn keskitytään tässä työssä kappaleessa 2.2.

Asioiden internetin yleistyessä on melko varmaa, että isoa osaa tuotetusta raa'asta datasta ei tulla koskaan tarkastelemaan ihmissilmin sen jättimäisen määrän vuoksi, vaan tekoälyalgoritmit lukevat, käsittelevät ja visualisoivat sitä ihmistä varten. Erityisesti sosiaalisen median synnyttämän tiedon kohdalla massadatan vahvuus korostuu perinteisiin tiedonkeräysmenetelmiin verrattuna. Kuten yllä mainittiin perinteiset menetelmät kuten haastattelut ja lomakkeet, sekä perinteiset aineistot kuten valtioiden itse keräämät väestönlaskennat ja erilaiset rekisterit ovat hitaita ja kalliita tuottaa. Esimerkiksi Yhdysvalloissa väestönlaskenta suoritetaan vain kerran 10 vuodessa, jonka myötä ”voimassa oleva” aineisto vanhentuu nopeasti (Silva et al. 2014). Perinteisesti kerätyllä, “harvalla datalla”, on tosin vahvuutensa: sosiaalisesta mediasta louhittu sisältö on sattumanvaraista eikä sitä ole standardoitu. Millerin ja Goodchildin (2015) mukaan tästä tiedonkeräysmenetelmien vastakkainasettelusta ja tehokkaista tekoäly-avusteisista analyysimenetelmistä voi syntyä kokonaan uusi maantieteellisen tutkimuksen paradigma: ‘*data-driven geography*’, vapaasti suomennettuna ‘aineistolähtöinen maantiede’, jossa perinteinen tapa tehdä maantieteellistä tutkimusta muuttuu. Esimerkiksi massadatan myötä aineistona ei välttämättä ole satunnaisotanta vaan koko populaatio ja massadatan “sotkuisen” sisällön esikäsittely tai sen tekemättä jättäminen määrittelee pitkälti sen mihin aineistolla voi vastata. Lisäksi suurien tietomäärien myötä perinteinen kysymyksen asettelu saattaa tapahtua vasta kun tutkijalla on käsitys siitä, mihin kysymyksiin aineistolla voi ylipäätään vastata tai mitä kysymyksiä aineisto nostattaa (Miller & Goodchild 2015). Määrän ja vauhdin lisäksi massadata on myös monipuolista. Se voi olla tekstiä, kuvia, videoita ja ääntä, jotka ovat eri kielillä, eri aiheista ja eri painotuksilla. Suuresta tietomäärästä, sen vauhdista ja monipuolisuudesta johtuen

tiede- ja yritysmaailman kiinnostus massadataa kohden on suuri. Aiemmin mainittu massadatan hyödyntäminen tuotekehityksessä ja markkinoinnissa lienee itsestään selvää, mutta massadata on varmasti yksi syy niin ”digitaalisten ihmistieteiden” suosion kasvuun. Tässä tieteenlajissa tietojenkäsittelytiede, digitaaliset teknologiat ja ihmistiede kohtaavat, sekä siinä pohditaan näiden teknologioiden vaikutusta kulttuuriin ja instituutioihin (Terras 2011).

1.3.1 Instagram ja muut sosiaaliset mediat

Sosiaalisen median palveluja on lukuisia ja niistä eräs suosituimmista on Instagram. Instagram on vuonna 2010 perustettu pääosin älypuhelimien kautta käytettävä kuvapalvelu, jossa käyttäjät voivat julkaista kuvia ja videoita varustettuna kuvatekstein ja aihetunnistein. Instagramiin ladattava ja siellä julkaistava kuvasisältö on standardoitu tietyn kokoiseksi: alun perin kuvien standardikoko oli 612 pikseliä kertaa 612 pikseliä. Teräväpiirtoteknologioiden yleistyttyä älypuhelimissa ja tablet-tietokoneissa kuvakokoa kasvatettiin 1080 pikseliin vuoden 2015 kesäkuussa. Alun perin Instagram kehitettiin Applen valmistamille iPhone- ja iPad-tuotteille, mutta suosion kasvun myötä se julkaistiin myös Android-puhelimille vuonna 2012. Vuonna 2010 Instagram saavutti miljoona käyttäjää, mutta vuoden 2014 lopussa Instagramissa oli jo reilu 300 miljoonaa käyttäjää ja vuonna 2016 käyttäjiä oli 600 miljoonaa (Wagner 2016), joka tekee palvelusta yhden suosituimmista sosiaalisen median palveluista ja se nousikin vuonna 2016 toiseksi suosituimmaksi sosiaalisen median alustaksi ohitse Pinterestin (Greenwood et al. 2016). Vuoden 2018 kesällä mitatuissa käyttäjätilastoissa Instagramilla oli jo miljardi käyttäjää (Statista 2018a). Instagramin käyttäjistä puolet on 18-29 -vuotiaita, joka on tuplasti enemmän kuin 30-49 -vuotiaita ja Instagramin käyttäjistä naisia on hieman alle 60 % (Greenwood et al. 2016). Ylipäätään, Facebookia lukuunottamatta, sosiaalisen median käyttäjät ovat pääsääntöisesti nuorehkoja ja teknologian kanssa lapsesta asti tekemisessä olleita henkilöitä vaikka käyttäjäprofiili vaihteleekin sosiaalisen median alustojen välillä (Leung 2013; Greenwood et al. 2016).

Verrattuna muihin kuvapalveluihin, kuten Flickr:iin ja Facebookiin, Instagram erottuu vahvaan visuaalisuuteen pohjustavalla kuvallisella sisällöllään ja melko yleisellä julkaistujen kuvien ”geoleimauksella” (engl. *geo-tagging*). Kuvien geoleimaus on kuitenkin Instagramin historian aikana muuttunut valinnaisempaan ja hieman epätarkempaan suuntaan. Instagram-sovelluksen julkaisutyökalussa ei enää ole

tarkan sijainnin jakamisen mahdollisuutta, vaan sijainti merkitään mielenkiintoisten kohdepisteiden kautta, kun taas aikaisemmin käyttäjä pystyi jakamaan tarkan sijaintinsa luomalla tarkalle sijainnilleen oman kohdepisteensä (Hochmair et al. 2018). Tarkan sijainnin jakaminen on ilmeisesti poistunut Instagramista samanaikaisesti ”Photo map”-ominaisuuden, joka mahdollisti kuvien katselun kartalta, kanssa syksyllä 2016 (Cvetojevic et al. 2016; Newton 2016; Hochmair et al. 2018). Tarkan sijainnin lisäämisen mahdollisuuden poistuminen ei vaikuta tähän työhön lainkaan, sillä aineisto on kerätty aikana, jolloin kyseinen ominaisuus oli vielä käytössä. Tosin tulee huomioida, että todennäköisesti suurin osa geoleimatuista julkaisuista on sidottu jo olemassa olleeseen pisteeseen sen sijaan, että käyttäjä olisi luonut täysin oman pisteensä jokaiselle julkaisulleen. Nykyisin Instagramissa tarjolla olevat kohdepisteet vastaavat pitkälti Facebookin kohdepisteitä, sillä Facebook omistaa Instagramin ja on siivonnut Instagramin kohdepistetietokantaa yhdenmukaistaen sitä Facebook Places -kohdepistetietokannan kanssa (Hochmair et al. 2018). Kuvien lisäksi Instagramissa voi nykyään julkaista myös videoita, sekä itsestään poistuvia videoiden ja kuvien sarjoja eli “Instagram-tarinoita”, myös näihin on mahdollista lisätä sijaintitieto.

Verrattuna esimerkiksi Google Earthin käyttäjäsisällön kaltaiseen, ennen kaikkea elinympäristön dokumentointiin tähtäävään, sisältöön, Instagram-julkaisut ja muiden sosiaalisten medioiden julkaisut painottuvat käyttäjien kokemuksiin, tuntemuksiin ja visuaaliseen itseilmaisuun (Hochman & Manovich 2013; Silva et al. 2014; Redi et al. 2016; Sheldon & Bryant 2016). Toisin sanoen, Google Earth ja siihen sisältyvät katunäkymät ja panoraamakuvat pyrkivät viestimään miltä jokin paikka näyttää objektiivisemmalla tasolla, kun taas Instagram-julkaisut viestivät mitä palvelun käyttäjät kokevat, ajattelevat ja tuntevat jonain aikana ja jossain paikassa. Tätä kokemuspohjaisuutta alleviivaa Instagram-palvelun tarjoamat lukuisat valokuvien muokkausmahdollisuudet, joiden avulla käyttäjät voivat muokata kuvia vastaamaan paremmin heidän omia tuntemuksiaan ja kokemuksiaan valokuvanottohetken tunnelmasta ja mielialasta (Hochman & Manovich 2014). Nämä ominaisuudet tekevät Instagram-julkaisuista aineistona hedelmälliseltä vaikuttavan aineiston kvantitatiiviseen ja kvalitatiiviseen tutkimukseen, sekä näiden yhdistelmiin esimerkiksi suurien ihmisjoukkojen paikkakokemuksien spatio-temporaaliseen analyysiin.

1.4 Sosiaalinen media maantieteellisessä tutkimuksessa

Sosiaalisessa mediassa syntyvää tietoa on viime vuosien aikana pyritty hyödyntämään kasvavissa määrin tutkimuksessa, päätöksenteossa ja kaupankäynnissä (Batty 2010; Leung 2013; Stefanidis et al. 2013; Silva et al. 2014; Zhou et al. 2014; Lansley & Longley 2016; Hausmann et al. 2017, 2018; Martin & Schuurman 2017; Redi et al. 2017; Tenkanen et al. 2017; Fu et al. 2018; Hiippala et al. 2018). Satojen miljoonien käyttäjien kokoiset sosiaalisen median verkostot tarjoavat suuren määrän tietoa, jonka saaminen perinteisillä "käsimenetelmillä" olisi erittäin aikaa vievää ja kallista. Sosiaalisen median suuren suosion syiksi on arveltu olevan ihmisen erilaisten perustarpeiden tyydyttämisen (Leung 2013; Kellerman 2014; Sheldon & Bryant 2016) itsestään selvien syiden, kuten helpon yhteydenpidon ystäviin ja perheesen lisäksi. Näitä tarpeita ovat muun muassa itsensä toteuttaminen, itsetunnon parantaminen ja turvallisuuden hakeminen (Kellerman 2014). Esimerkiksi sisällön tuottaminen sosiaaliseen mediaan, kuten kuvan julkaiseminen, voi tyydyttää käyttäjän psyko-sosiaalisia tarpeita kuten huomion saamista, aggression purkamista, tunnustuksen saamista, viihtymistä ja kognitiivisten tarpeiden tyydytystä (Leung 2013). Osittain näiden piirteiden takia tieteellinen kiinnostus sosiaalista mediaa kohtaan on kasvanut (Martin & Schuurman 2017), sillä sosiaalisen median julkaisut ovat vapaaehtoisesti luovutettua tietoa käyttäjiensä ajatuksista, kokemuksista ja tuntemuksista. Koordinaattitietojen sisällyttäminen sosiaalisen median julkaisuihin, geoleimaaminen, tekee sosiaalisen median aineistoista tavallaan vapaaehtoisesti tuotettua paikkatietoa. Tähän perustuen sosiaalisen median eri alustojen tuottama massadata on mahdollistanut suurikokoisten laadullisten aineistojen hyödyntämisen, johon ei olla ennen juurikaan ryhdytty nopeiden analyysimenetelmien ja tarpeeksi suuren tietokonelaskentatehon puuttuessa (Martin & Schuurman 2017).

Sosiaalisen median paikkatietoja on tutkittu spatio-temporaalisesti eri mittakaavoista. On globaaleja ja useamman vuoden tarkasteluja, sekä selkeästi rajatumpia kaupunginosien tarkasteluja (Graham & Zook 2011; Stefanidis et al. 2013; Martin & Schuurman 2017; Fu et al. 2018; Hiippala et al. 2018) jopa tiettyihin juhlapyhiin kohdistuvia tarkasteluja (Hochman & Manovich 2013). Kaupunkien dynamiikkaa ja sosiaalista käyttäytymistä on tutkittu Instagram-kuvien sekä Foursquare-aineistojen avulla muun muassa suurien maailmankaupunkien välillä (Silva et al. 2014). Kyseisessä tutkimuksessa luotiin ihmismassojen liikkumisesta "kaupunkikuvia", joissa eri aktiviteetteja ja niiden välillä tapahtuvia siirtymisiä visualisoitiin matriisiin, jolloin

kaupunkien väliset erot ihmismassojen vapaa-ajan tottumuksissa visualisoituivat yllättävän selkeästi. Silva ja kumppanit (2014) saivat näyttöä myös siitä, ettei sosiaalisen median alustalla ole vaikutusta julkaisuaktiivisuuteen, tosin tässä tapahtunee muutoksia eri alustojen suosioiden kasvaessa ja laskiessa ja esimerkiksi alustojen suosiossa valtioiden välillä voi olla suuriakin eroja kulttuuristen tai poliittisten syiden vuoksi. Instagramin lisäksi myös Flickr on kuviin keskittyvä sosiaalisen median palvelu, jota on käytetty aineistona tutkimuksessa (Crandall et al. 2009). Flickr on kuitenkin käyttäjämäärissä mitattuna selkeästi pienempi sosiaalisen median alusta verrattuna Instagramiin (Statista 2018a).

Tekoälyä ja koneoppimista hyödyntävien menetelmien kehittyminen ja yleistyminen on mahdollistanut laajojen sosiaalisen median aineistojen tutkimuksen. Esimerkiksi tekstiä käsittelemällä kieliteknologisin menetelmin siinä ilmeneviä aiheita ja sävyjä voidaan mallintaa ja visualisoida (Crandall et al 2009; Lansley & Longley 2016; Martin & Schuurman 2017; Fu et al. 2018). Kuvajulkaisuja on analysoitu koneoppimismenetelmin esimerkiksi siten, että pelkästään kuvan avulla on selvitetty jonkin kohteen maantieteelliset koordinaatit ja siten lisätty se kartalle (Neuhold et al. 2017). Semanttista segmentointia käytetään hyödyksi esimerkiksi avoimessa kartta- ja katunäkymäpalvelussa, Mapillaryssa, jossa muun muassa liikennevalot ja -merkit lisätään automaattisesti kartalle tekoälyn tunnistettua ne sille syötetyistä kuvasarjoista (Neuhold et al. 2017).

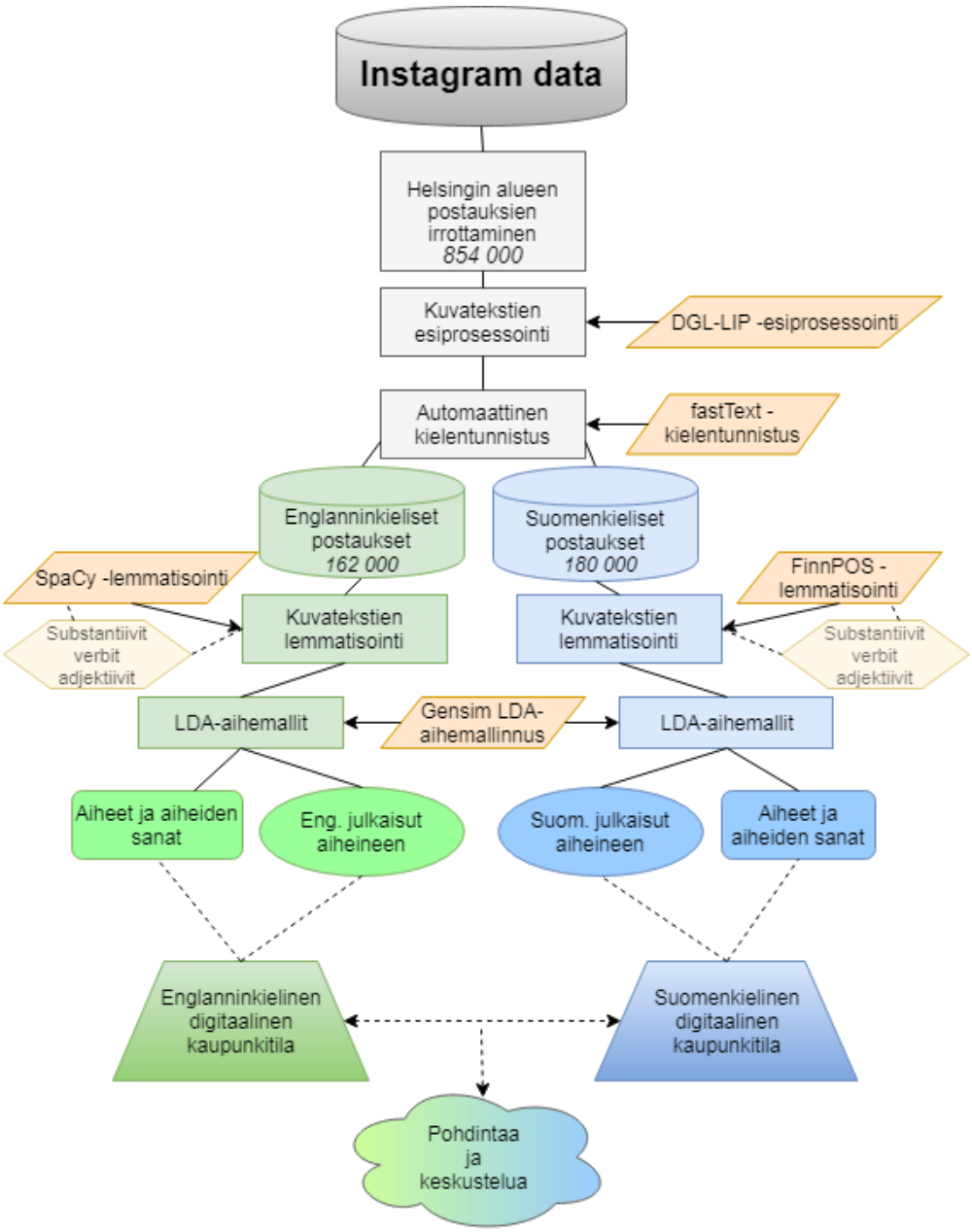
Pitkäaikaisten ja globaalien tarkastelujen lisäksi lyhyen aikavälin Instagram-käyttäytymistä on tutkittu muun muassa Tel Avivin kaupungissa Israelissa tarkastelemalla julkaisuaktiivisuuden muutoksia muutaman viikon aikana, jolloin Israelissa juhlitaan useampaa kansallista pyhäpäivää (Hochman & Manovich 2013). Hochmanin ja Manovichin tutkimuksessa julkaisutahdissa Tel Avivissa tapahtui merkittävä julkaisumäärien putoaminen kaatuneiden sotilaiden muistopäivänä, kun taas Israelin itsenäisyyspäivänä julkaisutahti kiihtyi jatkuvasti aamusta iltaan nousten todella korkeaksi verrattuna tavanomaiseen julkaisutahtiin. Lisäksi julkaisujen spatiaalisessa ja ajallisessa sijoittumisessa paljastui eroavaisuuksia käyttäjien välillä. Itsenäisyyspäivää seesteisemmin viettävät käyttäjät julkaisivat kuviaan aikaisemmin päivällä virallisen seremoniapaikan lähettyvillä, kun taas railakkaammin juhlivat käyttäjät julkaisivat vapaamuotoisemmasta juhlatapahtumasta myöhempään illalla. Vain harva käyttäjä oli julkaissut molemmissa tapahtumissa, josta

pääteltiin käyttäjäkuntien muodostavan toisistaan erillään olevat yhteisöt Tel Avivissa (Hochman & Manovich 2013). Tapahtumilla on todettu olevan vaikutuksia myös eri paikoista tapahtuvaan julkaisuaktiivisuuteen: esimerkiksi Belo Horizontessa jalkapallostadion korostui julkaisupaikkana vain niinä päivinä, kun siellä pelattiin jalkapallopeli (Silva et al. 2014).

Instagramin lisäksi Twitteristä louhittua aineistoa on käytetty muun muassa eri aiheista keskustelemisen spatiaaliseen mallintamiseen Lontoossa ja muualla (Jenkins et al. 2016; Lansley & Longley 2016; Fu et al. 2018). Käyttäjien sosioekonomisella taustalla, iällä ja etnisyydellä todettiin olevan vaikutus julkaistujen aiheiden valintaan, mutta myös fyysisen ympäristön todettiin vaikuttavan aiheisiin. Lontoolaisten Twitter-julkaisujen aiheet vaihtelivat suuresti alueelta toiselle ja esimerkiksi urheilustadioneilta lähetetyt Twitter-julkaisut erosivat aiheiltaan esimerkiksi ostoskeskuksista tai metroasemilta lähetetyistä (Lansley & Longley 2016). Vastaavanlaisia aihemallinnusta sosiaalisen median julkaisujen kanssa hyödyntäviä tutkimuksia on tehty myös kaupunginosittain (Martin & Schuurman 2017). Aihemallinnuksen lisäksi sosiaalisen median julkaisuvirtojen reaaliaikaisella tekoälyavusteisella louhinnalla on todettu olevan melko luotettavia eri käyttötarkoitusissa, esimerkiksi reaaliaikaisten liikenne- ja onnettomuustiedotuksien laadinnassa (Liu et al. 2014; Xu et al. 2015; Zhou & Zhang 2016).

2.0 Aineisto ja menetelmät

Tässä kappaleessa kuvaillaan käytössä olevaa aineistoa ja menetelmiä. Aineistona tässä työssä on Instagramista louhitut julkaisut, erityisesti niiden kuvatekstit, vuosilta 2015 ja 2016. Jotta aineistosta pystytään luomaan aihemalleja ja kuvatekstien analysointi olisi ylipäätään mahdollista, sitä täytyy esikäsitellä huomattavan paljon. Esikäsittelyyn sisältyy julkaisujen kuvatekstien muokkaaminen automaattiselle kielentunnistukselle sopivaan muotoon. Kun kuvatekstien kielet on tunnistettu, niistä valitaan suomen- ja englanninkielisiksi tarpeeksi suurella todennäköisyydellä tunnistetut kuvatekstit omiksi aineistoikseen. Suomen- ja englanninkielisten julkaisujen kuvatekstien sanat muunnetaan niiden perusmuotoon eli lemmatisoidaan, koska aihemallinnus tulkitsee saman sanan eri taivutusmuodot eri sanoiksi. Lemmatisoinnin jälkeen kuvateksteistä poistetaan informaatioltaan köyhät sanat eli niin sanotut hukkasanat (engl. *stop words*). Tämän jälkeen siirrytään varsinaiseen käsittelyyn, jossa tekstit muutetaan numeeriseen esitysmuotoon aihemallinnukselle, jotta sen käyttämät laskennalliset menetelmät olisivat mahdollisia. Aihemallinnus ajetaan halutuilla asetuksilla molemmille kieliaineistoille ja lopputuotteena syntyy taulukkoon koottu lista aiheista, aihekohtaisesti tärkeimpien sanojen ja koherenssipisteytyksen kera. Taulukon lisäksi aiheet aggregoidaan aihekennoihin spatiaalisen esittämisen mahdollistamiseksi, sekä aiheiden temporaalisia piirteitä esitellään kuvaajin. Aihemallinnus toteutetaan koko Helsingin mittakaavassa, mutta myös kolmeen kaupunginosaan keskittyen. Kaupunginosakohtaisessa aihemallinnuksessa aihemallit rakennetaan kaikkien lemmatisoitujen sanojen lisäksi myös pelkästään substantiiveja ja verbejä sekä pelkästään adjektiiveja käyttäen. Näistä tuloksista pyritään saamaan vastaus tutkimuskysymyksiin, jotka ovat esitelty johdannon alussa. Suurin osa työstä on tehty Python 3 -ohjelmointiympäristössä Pandas-, GeoPandas- ja Gensim-kirjastoilla pois lukien yksittäinen suomenkielisen aineiston esikäsittelyn vaihe ja paikkatietoanalyysit, jotka ovat tehty avoimilla paikkatieto-ohjelmistoilla: QGIS ja GeoDA. Tässä kappaleessa yleispiirteisesti kuvattua työnkulkua on havainnollistettu kuvassa 3 olevaan vuokaavioon.



Kuva 3. Vuokaavio tämän opinnäytetyön työnkulusta aineiston esikäsittelyyn, aihemallinnukseen ja tuloksiin keskittyen.

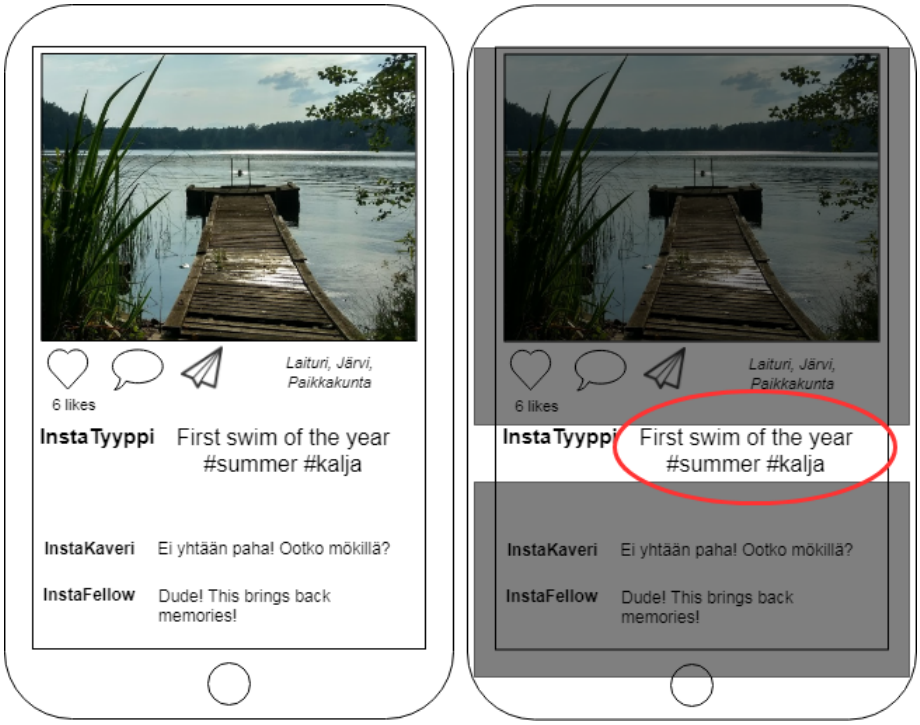
Eri kielellä tehtyjen julkaisujen välisten erojen löytämiseksi Helsingistä mallinnettavia aiheita tarkastellaan koko kaupungin mittakaavassa ja muutaman kohdealueen kautta. Kohdealueiksi tässä työssä valikoituivat Suomenlinna, Töölö ja Kallio. Töölö ja Kallio ovat Helsingin kaupunginosia, jotka ovat pääsääntöisesti asuinalueita, mutta niissä sijaitsee myös paljon kaupallista toimintaa sekä jonkin verran nähtävyyksiä. Suomenlinna on vastaavasti maailmanperintö- ja matkailukohde, mutta myös kaupunginosa, jossa asuu noin 800 asukasta. Kaikilla näillä alueilla tapahtuu Jan Gehlin hahmottelemia välttämättömiä, valinnaisia ja sosiaalisia aktiviteetteja (taulukko 1), koska kaikilla alueilla on asuntoja, työpaikkoja, palveluita ja vapaa-ajan

kohteita. Tosin eroja aktiviteeteissa on todennäköisesti sekä alueiden välillä, mutta myös alueiden sisällä kieltenvälisesti, joten voidaan olettaa tämän näkyvän myös mallinnettavista aiheista.

Töölö on vilkkain liikenteeltään ja kaupalliselta toiminnaltaan, johtuen Kampin keskuksen ja Rautatientorin läheisyydestä sekä vilkkaasti liikennöidyistä kaduista. Töölö siten muistuttaa muiden länsimaisten kaupunkien liikekeskustaa. Esimerkiksi Temppeliaukion kirkon ja Sibelius-monumentin myötä Töölön alueen sisällä on luultavasti näiden osalta suurta vaihtelua Instagram-julkaisujen aiheissa suomen- ja englanninkielisen aineiston välillä. Suomenkielisessä aineistossa on todennäköisempää, että kirkko ja monumentti eivät näy yhtä selkeästi, koska ne lienevät vain yksi arkielämän ympäristön piirre, kun taas englanninkielisissä julkaisuissa nähtävyydet voivat olla tärkeimpiä syitä käydä Töölön alueella. Kallio puolestaan on vilkkaudeltaan hieman rauhallisempi, sillä se on pääsääntöisesti asuinkerrostalojen aluetta ja kaupallinen toiminta rajoittuu katutason kivijalkaliikkeisiin, eikä aineiston aikajaksona Kalasatamaan syyskuussa 2018 avattu kauppakeskus Rediä vielä ollut olemassa rakennustyömaata lukuun ottamatta. Kaupallinen toiminta myös eronnee Töölön toiminnasta tuotteiden ja palvelujen hieman alhaisempina hintoina. Kallion alue koostuu käyttämässäni rajauksessa Harjun, Linjojen ja Torkkelinmäen pienalueista. Suomenlinna on näistä kahdesta selkeästi eroava alue, sillä se on saaressa sijaitseva UNESCOn maailmanperintökohteena toimiva linnoitus, ulkoilmamuseo ja pienimuotoinen asuinalue. Tämän tarkastelun puitteissa Suomenlinna on erityisen mielenkiintoinen, sillä suurin osa tarkastelun alla olevasta väestöstä nimenomaan vierailee siellä, jonka vuoksi oletettavasti suuria eroja Instagram-julkaisujen aiheissa ja sävyissä ei ilmene näiden ryhmien välillä.

2.1 Instagram-aineisto

Tämän työn pohja-aineistona toimii Helsingin Yliopiston Digital Geography Lab -tutkimusryhmän kehittämällä HUGOS-työkalulla (Tenkanen 2017) kerätty Instagram-aineisto, jonka ajallinen kattavuus tämän työn osalta alkaa vuoden 2015 alusta ja päättyy vuoden 2016 toukokuuhun. Aikaraamin pysähtyminen vuoden 2016 toukokuuhun johtuu Instagramin rajapintapalvelun käyttöoikeuksien muutoksista (Instagram for Developers 2016). Vuoden 2018 alussa paljastuneen Cambridge Analyticaan liittyvän skandaalin seurauksena Instagramin rajapintapalvelun avoimuutta ja toimintoja on rajoitettu entisestään.



Kuva 4. Havainnollistus Instagram-julkaisusta ja siitä julkaisun osasta (punaisella ympyröity), johon tässä työssä pääsääntöisesti keskitytään. Esimerkkijulkaisun kuva on tämän työn kirjoittajan itse ottama.

Aineisto on kerätty keväällä 2016 osana Helsingin yliopiston Social Media data for Conservation science -hanketta varten, joka alkoi maaliskuussa 2016 ja päättyy helmikuussa 2020. Rajapinnasta noudettu aineisto syötettiin PostGIS-tietokantaan, josta tämän työn aineisto valikoitiin paikkatietokyselyllä ja tallennettiin pistemuotoiseksi paikkatiedoksi avoimeen GeoPackage-tiedostomuotoon. Syy kyseisen tiedostomuodon käyttöön geoinformatiikassa yleisesti käytetyn Shapefile-tiedostomuodon sijasta juontaa juurensa shapefile -tiedostomuodon rajoittuneisuudesta suurten tiedostokokojen ja pitkien tekstien kanssa. Shapefile ei pysty esittämään yli 2 gigatavun kokoisia aineistoja, eikä tekstikentät voi olla 254 merkkiä pidempiä, mutta Instagram-julkaisujen kuvatekstit, joihin tässä työssä pääsääntöisesti keskitytään (kuva 4) voivat olla huomattavasti pidempiä. Tämä huomattiin käytännössä tässä työssä, sillä alun perin shapefileksi tallennetusta aineistosta hävisi suuri määrä ominaisuustietoja ja pitkät kuvatekstit loppuivat kesken.

Käytössä oleva Instagram-aineisto on rajattu Helsinkiin koko metropolialueen kattavasta aineistosta, koska Helsinki sisälsi selkeän ja silminnähtävän enemmistön koko aineistosta. Valitsemisessa käytettiin noin 100 metrin puskurivyöhykettä, sillä yllämainitun sijaintien kohdepistesitoutuneisuuden vuoksi Helsingin rajojen tuntumassa voi olla vahingossa esimerkiksi Vantaan puolelle lisätty kohdepiste, jonka tulisi olla Helsingin

puolella. Valikoitunut aineisto kattaa kaikki geoleimatut Instagram-julkaisut Helsingistä vuoden 2015 alusta vuoden 2016 maaliskuun viimeiseen päivään, jotka olivat keräyspäivänä avoimesti saatavilla. Osa aineistossa näkyvistä julkaisuista on sittemmin voitu poistaa palvelusta, osa käyttäjistä voi myös olla poistanut käyttäjäprofiilinsa tai asettanut profiilinsa näkyvyyden yksityiseksi keräämisen jälkeen. Näillä mainituilla asioilla ei ole vaikutusta tähän analyysiin. Geometrialtaan aineisto koostuu piste-muotoisista kohteista, joiden ominaisuustietoina on Instagram-julkaisun tietoja kuten kuvateksti, aihetunnisteet (engl. *hashtag*) ja kommenttien lukumäärä.

2.1.1 Aineiston kuvaus

2.1.1.1 Käyttäjät, aihetunnisteet ja kohdepisteiden sijainnit

Helsingistä kerätty aineisto koostuu noin 854 000 Instagram-julkaisusta, jotka ovat 152 267 eri käyttäjältä. Käsittelemätöntä aineistoa on visualisoitu kartalle kuvissa 11 ja 12. Selkeä enemmistö julkaisuista sijoittuvat Helsingin keskustan alueelle ja Helsingin ulkopuoliset keskittymät seurailevat melko vahvasti Helsingin sisäisiä pienempiä ala-keskuksia. Aineistossa on ns. superkäyttäjiä eli käyttäjiä, jotka ovat julkaisseet satoja Instagram-julkaisuja aineiston kuvaaman aikajakson aikana, mutta selvällä enemmistöllä on vain muutama julkaisu. Superkäyttäjiksi voisi aineistosta kuvan 5 perusteella luokitella kaikki, joilla on yli 250 julkaisua, sillä tämän jälkeen julkaisujen määrä per käyttäjä lähtee eksponentiaaliseen nousuun. Superkäyttäjiä on siis 98 kappaletta koko aineistossa, jolloin heidän vaikutus analyysissä jäänee lähes olemattomaksi, sillä heidän määränsä on häviävän pieni verrattuna reiluun 150 000 uniikin käyttäjän määrään. Käyttäjiä, joilla on enemmän kuin neljä julkaisua aineistossa on hieman alle 24 % koko aineistosta, vain yhden julkaisun käyttäjiä on peräti 45% koko aineistosta. Aineistossa ei vaikuta olevan juurikaan sisältöä automaattisesti luovia botteja, jotka ovat varsin yleisiä esimerkiksi Twitterissä (Varol et al. 2017). Botit näkyisivät selkeinä piikkeinä julkaisumäärissä, mutta superkäyttäjien joukossa tällaisia piikkejä ei ole (kuva 5). Kenelläkään käyttäjällä ei ole aineistossa yli 1000 julkaisua, suurin käyttäjäkohtainen julkaisumäärä on 924. Yli 500 julkaisun superkäyttäjiä on vain 17 kappaletta koko aineistossa ja yli 100 julkaisun käyttäjiä on 682. Tämän jälkeen julkaisujen määrän pieneneminen käyttäjää kohden pienenee tasaisesti loiventuen selkeästi kohti alle 5 julkaisun määriä. 20 suurimmasta superkäyttäjästä vain yksi on yritys, eräs kukkakauppa. Yrityksien ja

erilaisten organisaatioiden vaikutuksen minimoimiseksi analysoitavaan aineistoon valikoidaan vain sellaiset käyttäjät, joiden käyttäjänimissä ei esiinny yrityksiin tai muihin ei-henkilöihin viittaavia termejä. Näitä termejä ovat esimerkiksi: .fi, .com, shop, yhdistys, store ja kauppa -termit. Julkaisuja tällaisilta käyttäjiltä aineistossa on hieman yli 25 000. Menetelmä yrityksien poistamiseen on varsin mekanistinen, mutta parantanee aihemallinuksen onnistumista kuitenkin jonkin verran siten, että yksityishenkilöiden edustus aineistossa paranee.

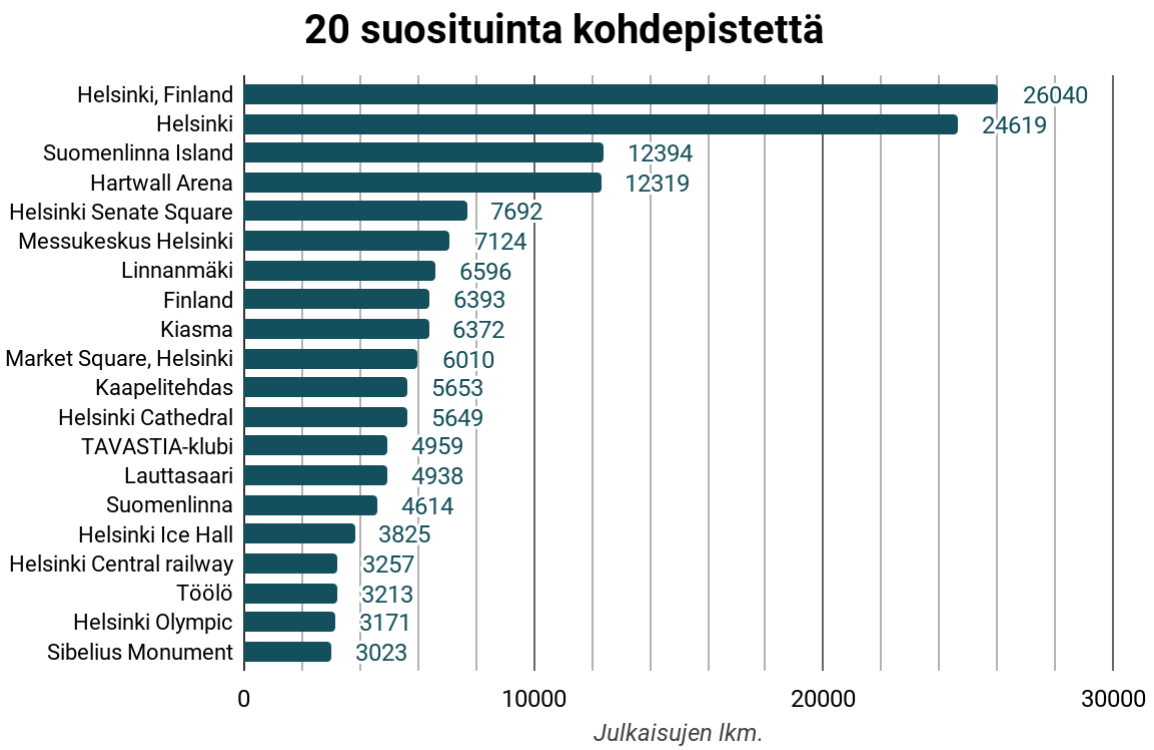
Kuva 5. Kuvaaja 500 eniten julkaisseista käyttäjistä eli superkäyttäjistä. Aineistossa on vain noin sata henkilöä, jotka ovat julkaisseet 2015-2016 välisenä aikana 250 kertaa.

Kuva 6. Käytetyimmät aihetunnisteet suomen- ja englanninkielisestä Instagram-aineistossa.

Aineistossa käytetyt aihetunnisteet, voivat toimia eräänlaisena ennakkotietona mahdollisista julkaisujen aiheista ja niiden käyttöä aihemallinnusta ohjaavana piirteenä onkin kokeiltu (Hong & Davison 2010). Kuvasta 6 on nähtävissä varsin selvästi, että neljä yleisintä aihetunnistetta liittyvät joko Suomeen tai Helsinkiin suoraan. Näiden lisäksi 19. yleisin aihetunniste on kyrillisin aakkosin kirjoitettu “Helsinki”. Tämä viestii venäjän olevan yksi yleisimmistä Instagram-kielistä Helsingissä, erityisesti sen vuoksi, ettei 25 yleisimmässä aihetunnisteessa esiinny muita kieliä englannin, suomen ja venäjän lisäksi. Ensimmäinen ruotsiksi kirjoitettu aihetunniste on “Helsingfors” sijalla 48. Yleisimmistä aihetunnisteista muut kuin maantieteellisiin sijainteihin liittyvät käsittelevät vuodenaikoja, matkustusta, tunnetiloja, vuorokaudenaikoja sekä luontoa. Ei ole järin yllättävää, että Helsinkiin sidotuista Instagram-julkaisuista paljastuisi aihemallinuksen jälkeen yhtenä suurena aiheena Helsinki. Yleisimmissä aihetunnisteissa ei varsinaisesti ole negatiivissävytteisiä aihetunnisteita, mutta selkeästi positiivissävytteisiä on kuvan 6 listauksessakin muutama (mm. love, happy ja friends), joka kertoo käyttäjien julkaisujen käsittelevän todennäköisesti enemmän positiivisia ja neutraaleja asioita. Täten on odotettavissa, että aineistosta mallinnettavat aiheet ovat myös positiivisiin ja neutraaleihin asioihin liittyviä. Lisäksi aihetunnisteissa esiintyvät sanat “vsco” ja

“vscocam” liittyvät valokuvien editointiin erikoistuvaan mobiilisovellukseen, eivätkä täten liity suoraan kaupunkitilaan tai ympäristöön vaan kyseisen julkaisun kuvaan ja siinä käytettyihin tehosteisiin.

Aihetunnisteiden lisäksi julkaisuissa esiintyvät suosituimmat kohdepisteet (kuva 7) voivat syventää kuvasta 6 johdettua ennakkokäsitystä julkaisuista mahdollisesti löytyvistä aiheista. Yleisimmät mainitut paikannimet aineistossa ovat yllätyksettömästi Helsinki, Suomenlinna ja Hartwall Arena. Yleisimmistä kahdestakymmenestä mainitusta sijainnista peräti kuusi liittyvät paikkoihin, joissa järjestetään suuria tapahtumia: Hartwall Arena, Messukeskus, Kaapelitehdas, Tavastia-klubi, Helsingin jäähalli ja Olympiastadion. Tapahtumapaikkojen lisäksi vapaa-ajan kohteita on mainituimpien paikkojen listalla neljä: Senaatintori, Linnanmäki, Kiasma ja Kauppatori. Tämä kielii mahdollisesti Instagram-käyttäjien taipuvuudesta tehdä julkaisuja normaalista arjesta poikkeavina (kenties positiivisina) hetkinä. Voikin olla, että Instagram-aineistot sopivat melko hyvin erilaisten tapahtumien tarkasteluun, kuten myös muualla on todettu (Hochman & Manovich 2013).



Kuva 7. Käsittelemättömän aineiston 15 yleisimmin mainittua sijaintia.

Toisaalta kuvasta 7 on nähtävissä myös aineiston kohdepisteisiin perustuvan spatiaalisen rakenteen ongelman. Yleispiirteisesti nimettyihin kohdepisteisiin, kuten “Helsinki” ja “Helsinki, Finland”, on sidottu

erittäin suuri määrä julkaisuja. Esimerkiksi Helsinki-sijaintiin sidottu julkaisu voi olla tehty mistä tahansa Helsingin alueelta. Lisäksi samaa paikkaa tarkoittavia kohdepisteitä on useampi: Helsinki ja Suomenlinna esiintyvät ainakin kahdessa eri kirjoitusasussa mainituissa kohdepisteissä (kuva 7). Kun selvittää, missä “Helsinki”-nimiseen kohdepisteeseen sidotut julkaisut sijaitsevat, paljastuu aineiston toinen ongelmallinen piirre: samalla nimellä esiintyvä piste voi sijaita useassa eri paikassa. Samanlainen ongelmallinen rakenne paljastuu tarkastellessa myös “Helsinki, Finland” -nimeen sidottuja julkaisuja sekä muita yleisimpiä kohdepisteitä. Syy tähän lienee käyttäjien itse tekemät sijainnit, jolloin samannimisiä pisteitä voi olla useassa paikassa.

Käyttäjät pystyivät jakamaan tarkkoja sijaintitietoja luomalla omia paikkoja Instagramiin, mutta omien paikkojen luomisen mahdollisuus poistettiin Instagramista vuoden 2016 aikana, kuten ylempänä mainittiin. Todennäköisesti omien paikkojen lisääminen päättyi Facebookin päätöksestä harmonisoida kohdepistetietokantansa Facebookin ja Instagramin välillä (Cvetojevic et al. 2016). Tämän vuoksi aineistossa on lukuisia käyttäjien itse lisäämiä vain kerran tai muutamia kertoja mainittuja sijainteja, joihin kuuluvat muun muassa “Huipulla”, “Kotona”, “Oma sänky” ja “Helvetti”. Näiden lisäksi “Koti” mainitaan sijaintina aineistossa noin 500 kertaa. Nämä erikoiset sijainnit eivät ole yllättäviä Helsinkiin keskittyvässä Instagram-aineistossa, sillä samankaltaisia geoleimauksia on todettu myös useasta muusta kaupungista ympäri maailmaa (Hochmair et al. 2017). Virallisten paikannimien lisäksi alueiden kutsumanimiä näkyy aineistossa, sekä sellaisia paikannimiä, joita Helsingistä kerätyssä aineistossa ei uskoisi olevan kuten Stockholm (22 kpl), New York (15 kpl) ja Kuopio (12 kpl). Nämä saattavat olla käyttäjien tahallaan lisäämiä vääriä paikannimiä, joilla kenties pyritään kommentoimaan miltä Helsinki näyttää tai tuntuu. Taustalla voi myös olla täysin vilpittömästi lisätty väärä paikannimi. Esimerkiksi turisti voi jostain syystä luulla olevansa Tukholmassa tai Kuopiossa Helsingin sijaan ja lisätä siten tämän nimisen paikan Instagramiin omaan sijaintiinsa ja geoleimaamalla julkaisunsa tähän luomaansa pisteeseen. Paikannimien lisäksi julkaisu sijaintimainintoja on luotu tapahtumien ympärille kuten Flow Festival (2253 kpl), Weekend Festival (1036 kpl) ja Naisten Kymppi (45 kpl).

Kohdepisteisiin perustuvan sijaintirakenteen kanssa on tärkeää huomioida, että maininta sijainnista julkaisussa ei välttämättä tarkoita sitä, että julkaisu käsittelee mainittua sijaintia tai että julkaisu tehdään sillä hetkellä kohdepisteen läheisyydessä. Kohdepistettä ei välttämättä ole luotu oikealle sijainnilleen, josta esimerkkinä toimii Helsingin Kauppatorilta löytyvä muutama julkaisu, joiden kohdepisteen nimi viittaa eri sijaintiin, kuin missä kohdepiste oikeasti sijaitsee: Tukholmaan ja Kuopioon. Mikään ei myöskään estä käyttäjää tekemästä Instagram-julkaista tapahtumasta jälkikäteen ja täysin eri sijainnista, jolloin esimerkiksi tammikuussa lisätty julkaisu voi liittyä useamman vuoden takaiseen kesätapahtumaan toisella paikkakunnalla. Tämä aineistopiirre vastaa erittäin hyvin Rosen (2017) ja Maleckin (2017) kuvailua digitaalisesta kaupunkitilasta paradoksaalisena: käyttäjä sijaitsee fyysisesti jossain, hän on yhteydessä digitaaliseen kaupunkitilaan fyysisesti olemassa olevan verkkoinfrastruktuurin kautta, hän surffaa verkossa joka on samanaikaisesti läsnä kaikkialla ja ei missään, sekä tekee Instagram-julkaisun edellisen vuoden Flow Festival -tapahtumasta kotoaan, joka sijaitsee esimerkiksi Oulussa.

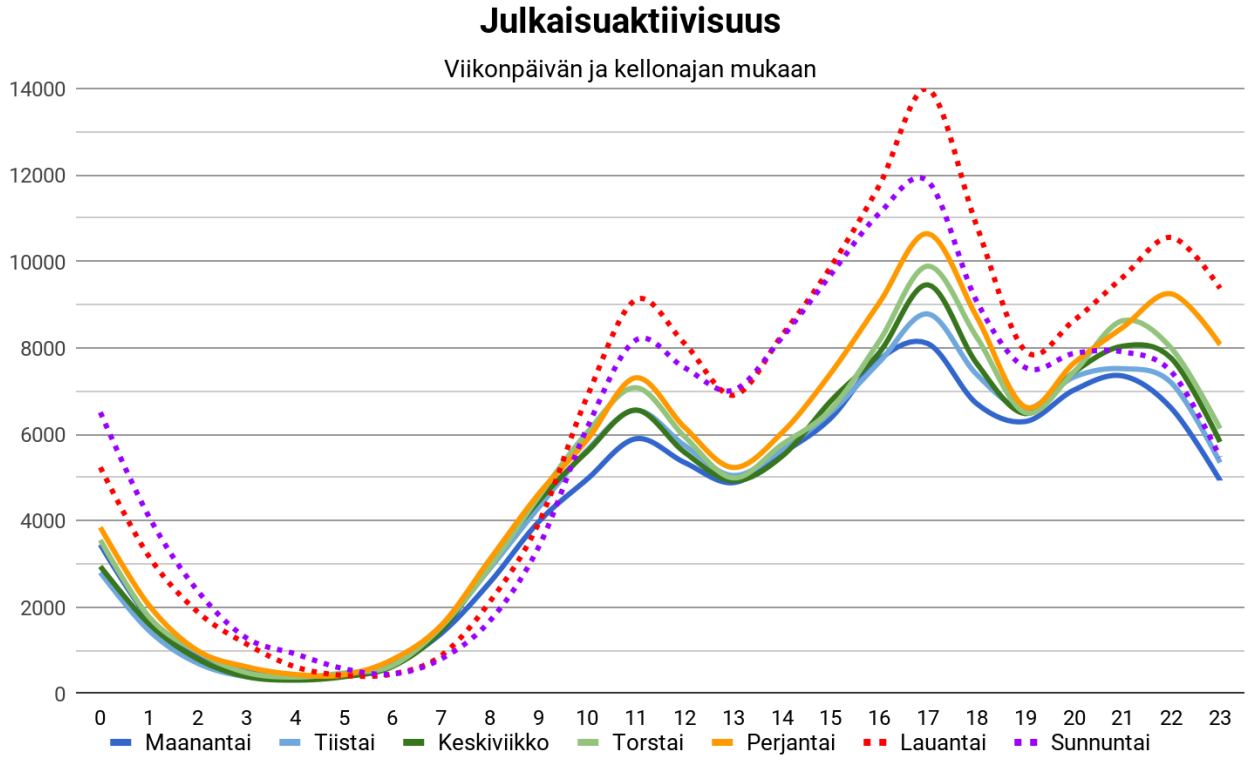
Paradoksaalinen luonne on osa sosiaalisen median aineistoja (Bendler et al. 2014), eikä sen aiheuttamaa vaikutusta ole tässä työssä siten pyritty minimoimaan. Paradoksaalisen luonteen lisäksi omien paikkojen luomisen mahdollisuus on keino, jolla käyttäjät pystyvät ottamaan haltuun digitaalista kaupunkitilaa ja antamaan sille omia merkityksiään, kuten saattaa olla esimerkiksi Helsingistä löytyvien New York -paikkamainintojen kanssa. Tosin uusia paikkoja ei voi lisätä Instagramiin enää, joten digitaalisen kaupunkitilan “haltuunottoa” ei pysty enää tällä tavoin tekemään. Tästä Instagram-aineiston paikkamainintojen ristiriitaisesta piirteestä huolimatta, huomattava enemmistö Instagram-julkaisuista on todettu tehtävän varsin lähellä mainittua sijaintia (Cvetojevic et al. 2016), joka on tärkeä tämän työn kannalta ja minimoi yllä mainittua epävarmuutta. Lisäksi anakronistiset julkaisut eivät välttämättä tee aineistosta epäluotettavampaa vaan paljastavat piirteitä käyttäjiensä ajatusmaailmasta, esimerkiksi keskellä talvea tehty kesään liittyvä julkaisu voi kertoa käyttäjän muistelevan sinä johonkin paikkaan liittyvää mennyttä hetkeä tai hänen toivovan kesän saapuvan nopeasti tai mahdollisimman hitaasti.

2.1.1.2 Ajallinen rakenne

Ajallinen tarkastelu on tärkeää tämänkaltaisessa työssä, sillä julkaisuaktiivisuuden huiput ja laaksot kertovat aineiston luonteesta ja palvelun käyttäjien aktiivisuusajoista. Instagram-alustan suosion kasvun ja laskun visualisoimisen ajassa lisäksi, aikaelementti on tärkeä tässä työssä, kun raa'asta aineistosta muodostetaan kokonaiskuvaa ja erityisesti työn tuloksena syntyviä aiheita tarkastellessa. Esimerkiksi osa aiheista saattavat olla vahvasti sidoksissa johonkin kellonaikaan, viikonpäivään, kuukauteen tai vuodenaikaan. Aikatarkastelu mahdollistaa teoriakirjallisuudessa mainitun kaupungin pulssin tarkastelun (Batty 2010).

Tarkastellessa Instagram-aineiston ajallista jakaumaa seuraavista kuvista 8-12, aineistosta paljastuu mielenkiintoisia piirteitä. Julkaisemisessa on selkeä päivittäinen ja viikoittainen rytmiikka, sekä pidemmän aikavälin tarkasteluilla on selkeästi nähtävissä julkaisujen määrän kasvu, joka kielii Instagramin suosion kasvusta sosiaalisen median alustana. Kuvan 8 kuvaaja näyttää Instagram-julkaisujen rytmiikkaa viikonpäivittäin, josta paljastuu mielenkiintoinen piirre: viikon alku on hiljaisinta Instagram-julkaisuaikaa, mutta aktiivisuus kasvaa tasaisesti perjantaihin asti, jonka jälkeen aktiivisuudessa saavutetaan selkeä huippu lauantaina ja hieman matalampi, mutta edelleen perjantaita suurempi, huippu sunnuntaisin. Aikaisemmin kuvista 6 ja 7 huomatun aihetunnisteiden ja kohdepisteiden vapaa-aikakytköksen lisäksi tämä ajallinen piirre vahvistaa käsitystä Instagram-aineistosta vapaa-ajan asioihin keskittyvänä sosiaalisen median alustana. Kun tähän viikonpäivittäiseen tarkasteluun tuo lisäulottuvuuden ja katselee julkaisuaktiivisuutta eri viikonpäivän ja kellonajan mukaan, aktiivisuudesta paljastuu jälleen uusi rytmiikan piirre, kuten kuvasta 9 on nähtävissä.

Kuva 8. Instagram-julkaisuaktiivisuus viikonpäivittäin.

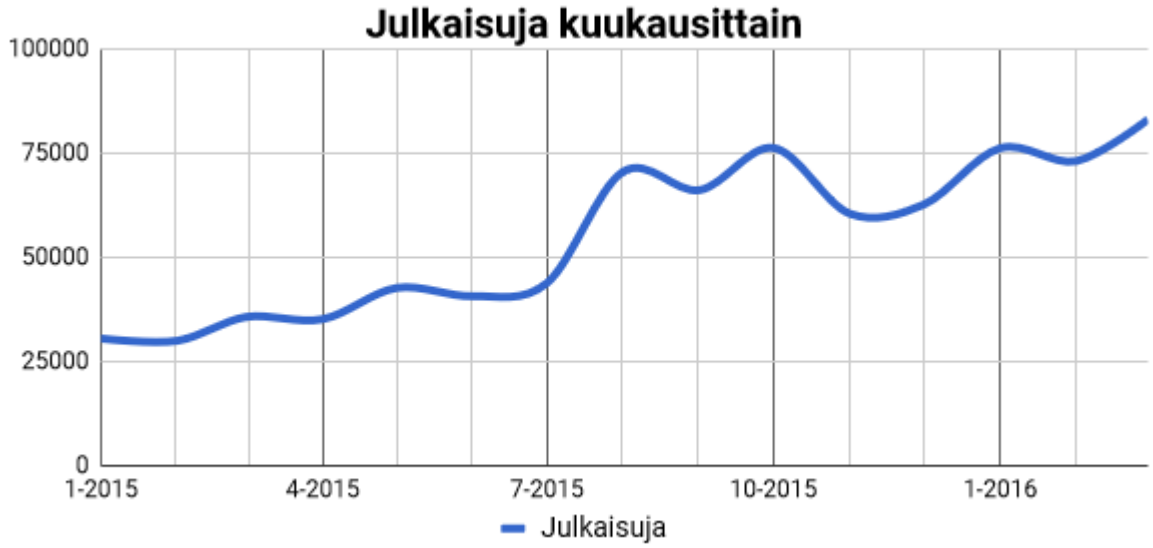


Kuva 9. Helsingistä kerätyn Instagram-aineiston ajallista jakautuminen eritelty viikonpäivän ja kellonajan mukaan, josta on selkeästi huomattavissa päivittäinen julkaisurytmiikka.

Kuvaa 9 tarkastelemalla vaikuttaa siltä, että kellonajan mukaan mitattuna julkaisuaktiivisuudessa on selkeä ja säännöllinen rytmi, joka toistuu samankaltaisena lähes jokaisena viikonpäivänä. Myöhäisen yön ja

aikaisen aamun tunteina julkaisuaktiivisuus on varsin yllätyksettömästi erittäin alhainen läpi viikon. Arkisin kello 6 ja viikonloppuisin kello 7 aktiivisuus alkaa kasvaa saavuttaen ensimmäisen huippunsa kello 11 aikoihin, josta se laskee lähes yhtä voimakkaasti kuin nousikin kello 13 asti. Kello 13 jälkeen aktiivisuus kasvaa ja saavuttaa päivittäisen huippunsa kello 17 aikaan, jonka jälkeen tapahtuu toinen notkahdus päätyen kello 19 aikoihin. Tästä aktiivisuus nousee vielä hieman muodostaen piikin kello 21 aikaan muina päivinä perjantaita ja lauantaita lukuun ottamatta, joiden korkeampi aktiivisuuspiikki saavutetaan kello 22 aikaan. Tämän jälkeen aktiivisuus tippuu alhaisimmalle tasolleen kello 3 ja 6 väliseksi ajaksi. Sunnuntai-illan vaikutus näkyy aktiivisuuden putoamisena, sillä on oletettavissa, että suuri osa käyttäjistä menee hieman aikaisemmin, jotta maanantaina jatkuvaan arkeen olisi saanut nukuttua tarpeeksi. Viikonlopun vaikutus näkyy myös aamuisin, jolloin julkaisuaktiivisuus on pienempää kuin arkena. Vapaa-ajan painottuminen julkaisuajoissa näyttäytyy aktiivisuudessa viikonlopun piikin lisäksi myös yleisinä ruokailuaikoina, klo 11, 17 ja 21-22. Tämä ei yllätä sillä, yksi käytetyimpiä aihetunnisteita on “food”, kuten kuvasta 5 on nähtävissä.

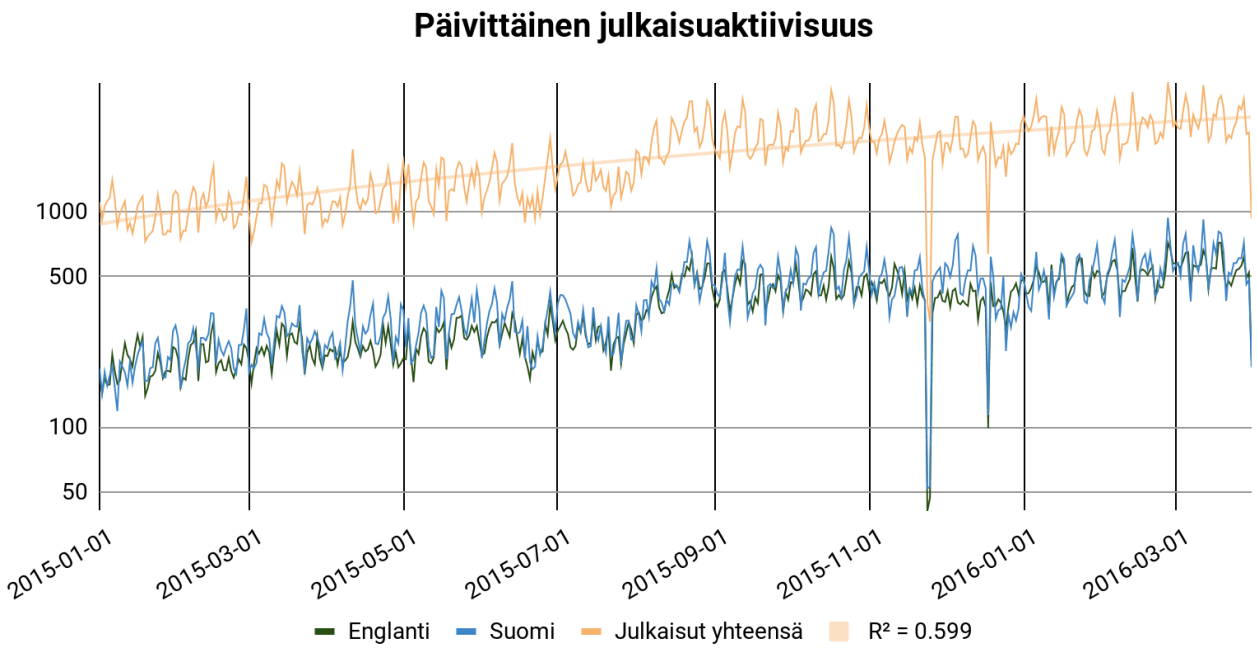
Instagram-kuvista suuren osan on todettu käsittelevän ruokaa, joten tämä ei sinänsä ole yllättävää, mutta havainto on mielenkiintoinen ja Instagram-aineistoja onkin hyödynnetty tutkittaessa käyttäjien ruokailutottumuksia eri puolilla maailmaa (Coary & Poor 2016; Chung et al. 2017). Kuvan 9 rytmiikkaa muistuttava rytmiikka on todettu myös Suomi24 -keskustelufoorumilla (Lagus et al. 2016). Julkaisuaktiivisuuden keskittyminen viikonloppuun ja perinteisten ruokailuaikojen yhteyteen entisestään vahvistaa käsitystä Instagram-alustasta vapaa-aikaan liittyvien julkaisujen sosiaalisena mediana, joka on huomattu myös muualla (Silva et al. 2014). Vaikka kellonaikakohtaisessa tarkastelussa paljastunut yleisten ruoka-aikojen suosio julkaisumäärien osalta on selkeästi havaittavissa, ajallinen rakenne vaikuttaa pääpiirteittäin seurailevan tavallista vuorokausirytmiä, joten Instagram-aineiston ajallinen luotettavuus vaikuttaa olevan varsin hyvällä tasolla, mikä ei ole itsestään selvää kaikkien sosiaalisen median aineistojen kanssa (Bendler et al. 2014).



Kuva 10. Instagram-julkaisujen määrän kehitys kuukausittain. Kehitys on nousujohteista aineiston ajallisen kattavuuden aikana, mutta Instagram-alustan suosio on jatkanut kasvuaan myös aikajakson jälkeen (Statista 2018a).

Pidemmän aikavälin julkaisuaktiivisuuden tarkastelussa on vähemmän mielenkiintoisia piirteitä. Kuukausittainen aktiivisuus (kuva 10) näyttää Instagramin suosion kasvun julkaisujen määrissä. Ero vuosien 2015 ja 2016 tammikuiden välillä on lähes kolminkertainen (ks. myös Hiippala et al. 2018). Kysymyksiä herättää vuoden 2015 heinä-elokuun vaihteessa tapahtuneen äkillisin käyttäjämäärän kasvu. Se näyttää ajoittuvan juurikin niihin aikoihin, kun Facebook osti Instagramin (Cvetojevic et al. 2016) ja on hyvin mahdollista, että noihin aikoihin useat Facebookin käyttäjät siirtyivät käyttämään myös Instagramia. Marras- ja joulukuun julkaisuaktiivisuudessa tapahtuu notkahdus, joka saattaa johtua vuodenaikaan osuvista juhlapyhistä. Tosin vapaa-aikaan vahvasti keskittyvänä alustana juhlapyhien luulisi pikemminkin lisäävän kuin laskevan julkaisutahtia, vaikkakin juhlapyhien on todettu tiputtavan julkaisuaktiivisuutta tiettyjen kansanryhmien osalta (Hochman & Manovich 2013; Manovich & Idaco 2017). Tämä notkahdus on erittäin selkeästi nähtävissä kuvasta 11 kahtena alaspäin suuntautuvana piikkinä sekä koko aineistossa, että suomen- ja englanninkielisiin julkaisuihin rajatuissa pienemmissä aineistoissa. Hieman kummastusta herättäen notkahduksien päivät ovat marraskuun 24. ja 25. päivä, sekä joulukuun 18. päivä. Aluksi erityisesti marraskuun 24. ja 25. päivien osalta tämä näytti aineistossa olevalta aikaleimavirheeltä, jolloin joulukuun 24. ja 25. päivä olisivat siirtyneet kuukauden verran taaksepäin, mutta notkahduspäivien kuvatekstejä tarkastellessa niistä ei paljastunut mitään ilmiselviä virheitä eikä anakronistisuuksia, mitkä viittaisivat aineiston aikaleimauksen epäonnistuneen. Tällaisia olisi olleet esimerkiksi kyseisten päivien julkaisujen

kuvatekstien keskittyminen jouluaattoon ja jouluun ylipäätään, mutta näin ei aineistossa ole käynyt. Jouluaaton ja joulupäivän päivämäärien kohdalla aktiivisuudessa ei tapahdu muista päivistä poikkeavaa notkahdusta. Sama notkahdus on myös havaittavissa käsittelemättömässä PostGIS-tietokannan aineistossa, johon aineisto on alun perin kerätty, joten kyseessä voi olla esimerkiksi huoltokatkosta johtuva aktiivisuuden tippuminen tai aineiston keräyksessä tapahtunut virhe, jolloin kyseisiltä päiviltä ole saatu kaikkia julkaisuja kerättyä. Tämä virhe-elementti ei vaikuttane työhön ja tuloksiin merkittävästi.



Kuva 11. Instagram-julkaisujen määrä päivittäin 2015-2016. Vuoden 2015 loppupuoliskolla näkyy selkeä putous julkaisuaktiivisuudessa, joka osuu marraskuun kohdalle. Kuvaajasta näkyy myös Instagramin suosion kasvu vuodesta 2015 vuoteen 2016.

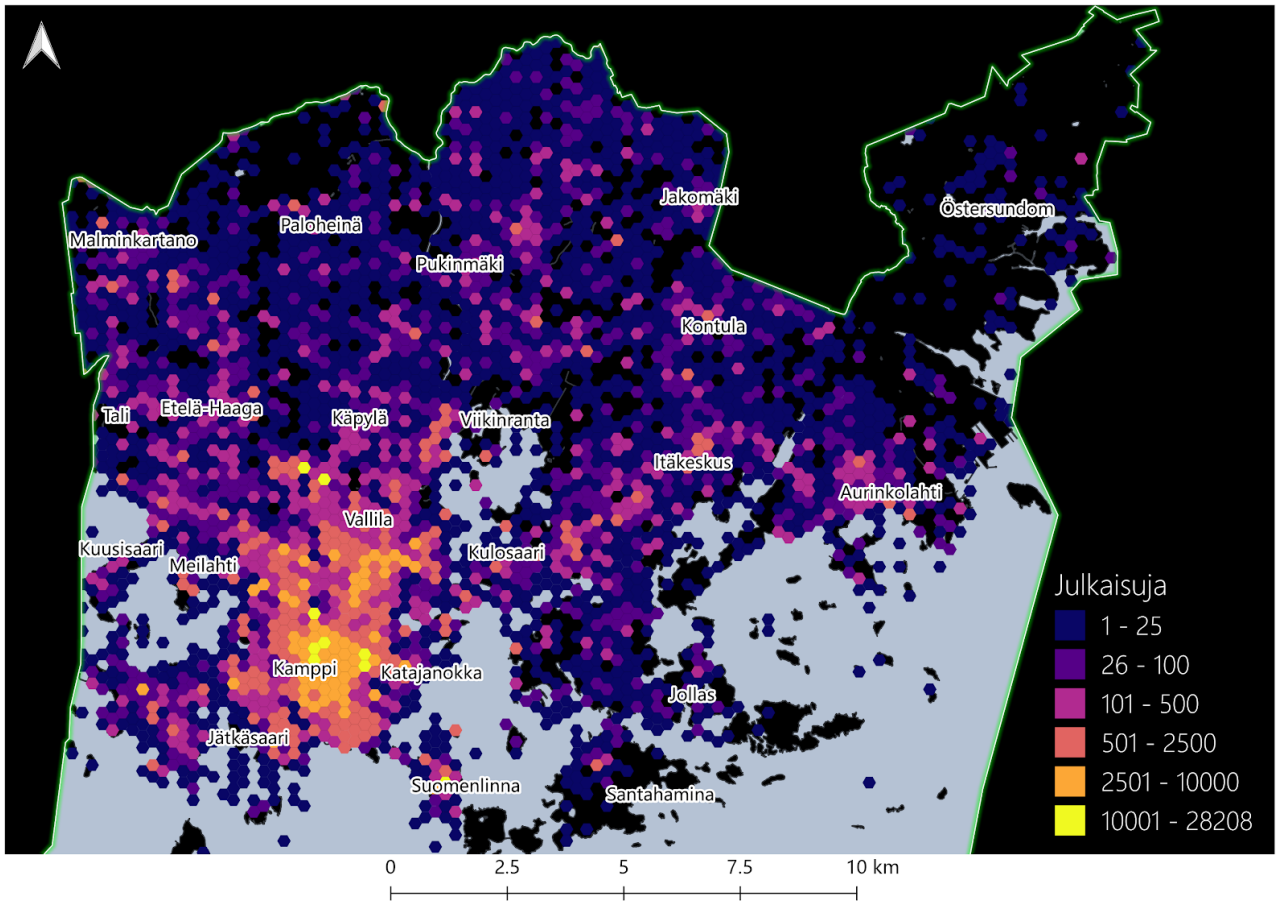
Julkaisun aiheen, ajan ja paikan kytkeytyminen todellisuuteen voi muodostua ongelmalliseksi tietyissä tapauksissa. Instagramissa julkaistava kuvat ei välttämättä ole julkaisuajankohtana otettu tai edes liity niihin koordinaatteihin, joissa käyttäjä julkaisuajankohtana on (Hochman & Manovich 2013; Jenkins et al. 2016, Hochmair et al. 2017). Tosin käyttäjä voi korjata tämän geoleimaamalla kuvan siihen liittyvään paikkaan myös jälkeenpäin. Nämä ongelmalliset piirteet ovat kuitenkin osa digitaalista kaupunkitilaa kuvaavissa aineistoissa, koska digitaalinen kaupunkitila on luonteeltaan paradoksaalinen ja monimuotoinen (Rose 2017; Kellerman 2014, 2015).

2.1.1.3 Alueellinen rakenne

Käsittelemättömästä aineistosta paljastuu varsin mielenkiintoisia piirteitä pelkästään visuaalisen tarkastelun myötä. Julkaisut näyttävät keskittyvän erityisesti alueille, joiden asukastiheys on korkea, kuten Helsingin keskustan alueelle ja pienempiin keskuksiin liikenteen solmukohtiin. Kun julkaisut aggregoidaan kuusikulmioista muodostuvaan 250 m x 250 m kennostoon (kuva 12), siitä on nähtävissä julkaisujen selkeä keskittyminen Helsingin keskustan alueelle. Kuusikulmioista koostuvan kennoston käyttäminen, perinteisemmän ruutupohjaisen ruudukon sijaan, johtuu kuusikulmaisen kennoston polygonien pienempään reunavaikutukseen, levinneisyyden todellisten muotojen selkeämpi välittyminen ja solujen topologisen naapuruuden yksinkertaisempi määrittely kuin neliskulmaisessa ruudukossa (Birch et al. 2007; ESRI 2018). Kennoja käytetään myös kaupunginosakohtaisessa aihevertailussa aihekennoina, joissa kenno luokittuu yleisimmän aiheen mukaan.

Kuten kuvassa 12 näkyvistä julkaisumääristä on nähtävissä, Helsingin ydinkeskusta muodostaa oman ison kokonaisuutensa ja sen pohjoispuolelle muodostuu hieman irrallinen Kallion kokonaisuus. Korkeimpien julkaisumäärien kennot ovat ydinkeskustan lisäksi Hartwall Arenan ja Suomenlinnan kohdilla. Pienempiä julkaisumäärien keskittymiä on siellä täällä, kuten Malmilla, Herttoniemessä, Itäkeskuksessa ja Lauttasaaressa. Kuvassa 12 julkaisut muodostavat suuren keskittymän ydinkeskustaan, josta lähtee ulospäin suuntautuvia sormimaisia spatiaalisia rakenteita. Selkeimmät sormet ovat Hämeentietä/Lahdenväylää ja Itäväylää seurailevat julkaisutihentymiä. Kuvissa 12 ja 13 näkyy myös muutamia julkaisuja vesialueiden päällä, jotka saattavat olla veneessä tai saaressa tehtyjä julkaisuja. Vesialueille sijoittuneet julkaisut voivat olla myös ”väärin” sijoittuneita käyttäjien luodessa kohdepisteitä. Kun käyttäjät vielä pystyivät luomaan omia paikkoja aineistoon, paikan tarkka sijainti tallennettiin joko älylaitteen GPS-laitteen avulla tai sovelluksen kartasta valitsemalla, jolloin paikannus- ja sijaintivirheet ovat mahdollisia. Esimerkiksi rannikolta tallennettu sijainti voi olla paikannusvirheen vuoksi siirtynyt jopa 20 metriä ja siten sijaita meressä. Kuvasta 12 ei suoraan näe, että julkaisuja olisi kertynyt mihinkään tiettyyn kohdepisteeseen suunnattomia määriä. Ylempänä kuvasta 7 kuitenkin voi huomata, että muutamien tietyn nimisten kohdepisteiden alle on sidottu suuret määrät julkaisuja. Tosin samalla nimellä esiintyy useita kohdepisteitä, jotka sijaitsevat useassa

paikassa eri puolilla Helsinkiä, joka vähentää tätä julkaisumäärien erityisen suurta kertymää yhteen tiettyyn kohdepisteeseen.



Kuva 12. 250 metrin kuusikulmaiseen kennostoon aggregoitu käsittelemätön Instagram-aineisto paljastaa julkaisujen spatiaalisen jakautumisen. Luokittelumenetelmä on luonnollinen hajonta, sillä se tuo aineiston sisäisen vaihtelun paremmin esille.

Vähäisten julkaisumäärien alueita ovat selkeästi Östersundom, Paloheinän pohjoispuoli, mutta myös osa saarista kuten Villinki, Vallisaari ja Pihlajasaari. Näistä saarista tosin Vallisaarta ei oltu vielä avattu yleisölle, vaan se oli Puolustusvoimien hallinnoima saari. Vallisaaren avaaminen kesällä 2016 näkyisi todennäköisesti yhtäkkisenä kesäisenä julkaisupiikkinä saaren alueelta. Saarien lisäksi mereisillä alueilla ei juurikaan ole julkaisuja, joka ei ole yllättävää Instagramin kohdepisteisiin perustuvan geoleimauslogiikan huomioon ottamisen jälkeen. Suurin aukko julkaisuissa sijaitsee Östersundomissa, joka ei myöskään yllätä, sillä alue on vahvasti metsä-, pelto- ja pientalovaltainen kaupunginosa. Östersundomin lisäksi useat luontoalueet kuten Keskuspuisto, Viikinranta, Kivinokka ja Uutela ovat alhaisten julkaisumäärien alueita. Pohjois-Helsingissä Paloheinän pohjoispuolella on nähtävissä selkeä aukko Keskuspuiston alueella julkaisujen alueellisessa rakenteessa kuvassa 12, joka kielii siitä, että Instagramin käyttäjät tekevät melko vähän julkaisuja

Keskuspuistossa verrattuna urbaaneihin alueisiin. Toinen selkeä aukko hieman etelämpänä on Viikinrannan alueella, jossa sijaitsee Viikin arboretum. Toisaalta aukot voivat olla näennäisiä, sillä samaan kohdepisteeseen voidaan liittää useita julkaisuja ja suurien viheralueiden, kuten Keskuspuiston, osalta Instagramissa voi olla vain yksi kohdepiste koko alueelle.

Taulukkoon 2 on laskettu tilastollisia muuttujia kuvan 12 kennostoon aggregoiduista julkaisuista, joista selviää, että kaikki Helsingin Instagram-julkaisut sisältyvät 3402 kennoon eli noin 84,5 neliökilometrin alueelle. Julkaisuja on keskimäärin noin 244 per kenno vaikkakin suurin yksittäinen joukko on yhden julkaisun kennot (469 kpl). Solujen mediaani on 15 ja hieman alle puolet kaikista kennoista pitääkin sisällään 10 tai pienemmän määrän julkaisuja. Keskihajonta on todella suuri, joka kertoo julkaisujen suuresta määrällisestä vaihtelusta soluittain.

Taulukko 2. Kuusikulmioiseen kennostoon (kuva 10) aggregoitujen julkaisumäärien tilastollisia muuttujia.

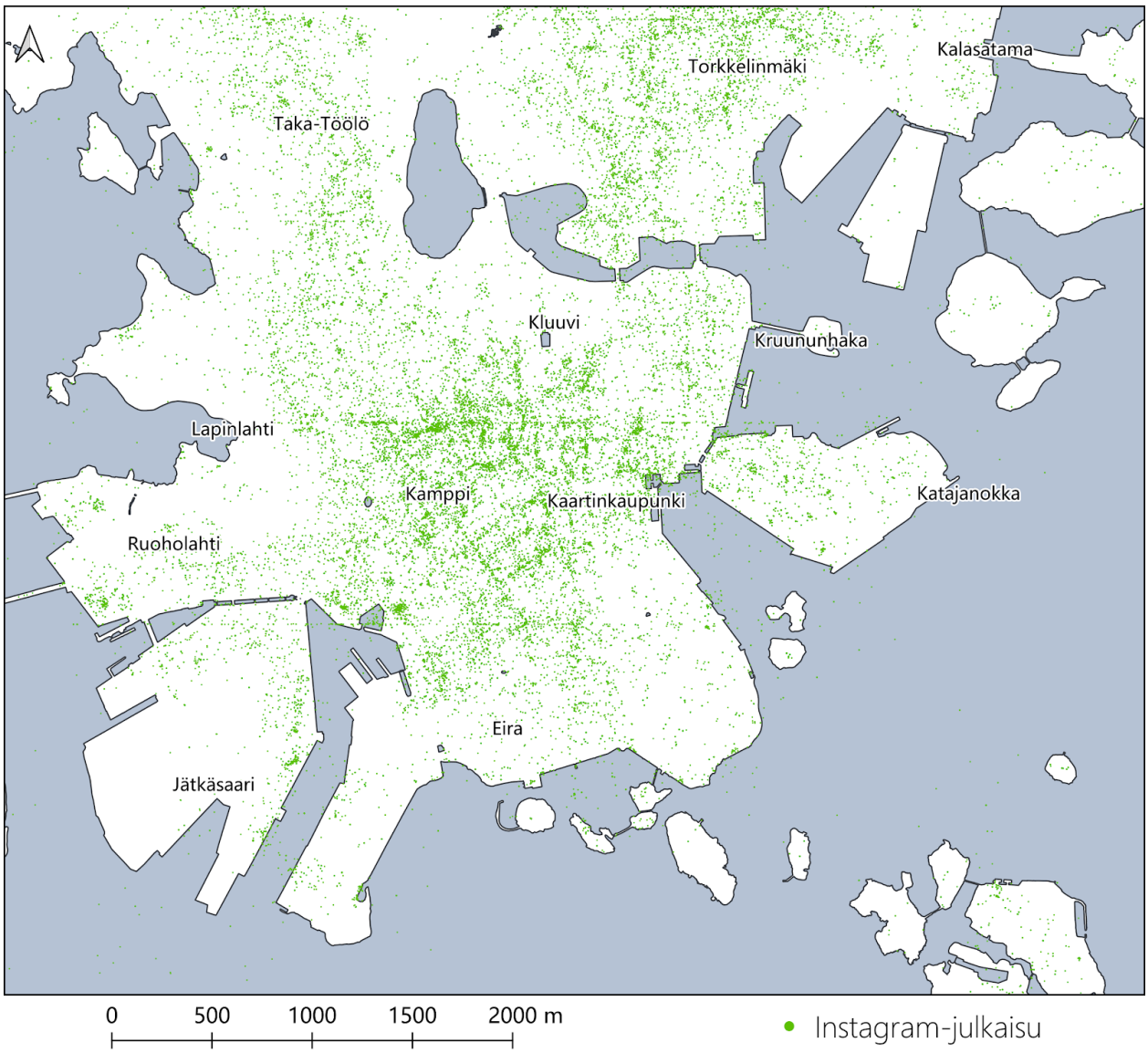
Lkm	Summa	Keskiarvo	Mediaani	Keskihajonta	Minimi	Maksimi
3402	828 579	243,56	15	1163,04	1	28208

Kohdepisteisiin spatiaalisesti perustuva aineistorakenne voi haitata aineiston ja tulosten analysointia jonkin verran, kuten ylempänä tässä tekstissä mainittiin. Kuvassa 7 näkyneet 20 suosituinta kohdepistettä ja erityisesti niistä suosituimmat voivat harmillisesti aiheuttaa pieniä ongelmia lopputulosten erittelyyn, sillä esimerkiksi “Helsinki” tai “Helsinki, Finland” nimisten kohdepisteiden sijainteihin on sidottu yhteensä reilu 50 000 julkaisuja. Toisaalta, ylempänä mainittu samannimisten kohdepisteiden sijainnillinen hajautuneisuus ei tee kohdepisterakenteesta niin suurta ongelmakohtaa, kuin mitä se voisi olla. Kohdepisteiden yleispiirteiset nimet, kuten “Helsinki”, tuovat kyseisiin julkaisuihin sijainnillisen virhetekijän mukaan myös siinä mielessä, että esimerkiksi kyseiseen kohdepisteeseen sidottujen julkaisujen todellinen sijainti voi olla missä päin Helsinkiä tahansa. On julkaisun lisänneestä käyttäjästä kiinni, mitä hän pitää Helsinkiä hyvin kuvaavana kohdepisteenä, jonka lisäksi kyseiset julkaisut eivät välttämättä lähellä sitä sijaintia, missä julkaisun kohdepiste sijaitsee. Tämä voi hieman vaikeuttaa aiheiden spatiaalista tarkastelua ja heikentää eri alueille syntyvää “aiheprofiilia”. Hyvä esimerkki tästä kohdepisterakenteesta on Korkeasaari, joka näyttää

kuvan 13 kartan oikeassa ylänurkassa vain muutaman julkaisun paikalta, mutta alueen pisteisiin on sidottu lähes 4000 Instagram-julkaisua.

Kuten ylempänä mainittiin, mikäli aineiston ajallinen kattavuus kattaisi vuoden 2016 kokonaan, Suomenlinnan kaakkoispuolella sijaitseva Vallisaari olisi varmasti kerännyt Instagram-julkaisuja. Kuvassa 12 Vallisaari on alue, jolta ei ole tehty lainkaan julkaisuja. Vallisaari avattiin yleisölle kesällä 2016 samantapaisena ulkoilmamuseona ja retkikohteena kuin viereinen Suomenlinnan saariryhmä, joka puolestaan on varsin suosittu Instagram-julkaisujen alue. Ennen avaamistaan Vallisaari oli ollut Puolustusvoimien hallinnoima saari, jonne oli pääsy kielletty ilman Puolustusvoimien hyväksyntää. Sosiaalisen median julkaisujen alueelliset rakenteet reagoivat siis varsin nopeasti oikeassa fyysisessä maailmassa tapahtuviin muutoksiin ja täten melko varmasti myös heijastavat fyysisessä kaupunkitilassa tapahtuvia muutoksia. Esimerkiksi vuoden 2018 kesällä Helsingin keskustaan avattiin uusi taidemuseo, Amos Rex, jonka näkyvyys sosiaalisen median julkaisuissa nousi varmasti avajaispäivänä ja -viikolla. Tämä piirre vahvistaa aineiston sopivuutta digitaalisen kaupunkitilan tarkasteluun ja analysointiin.

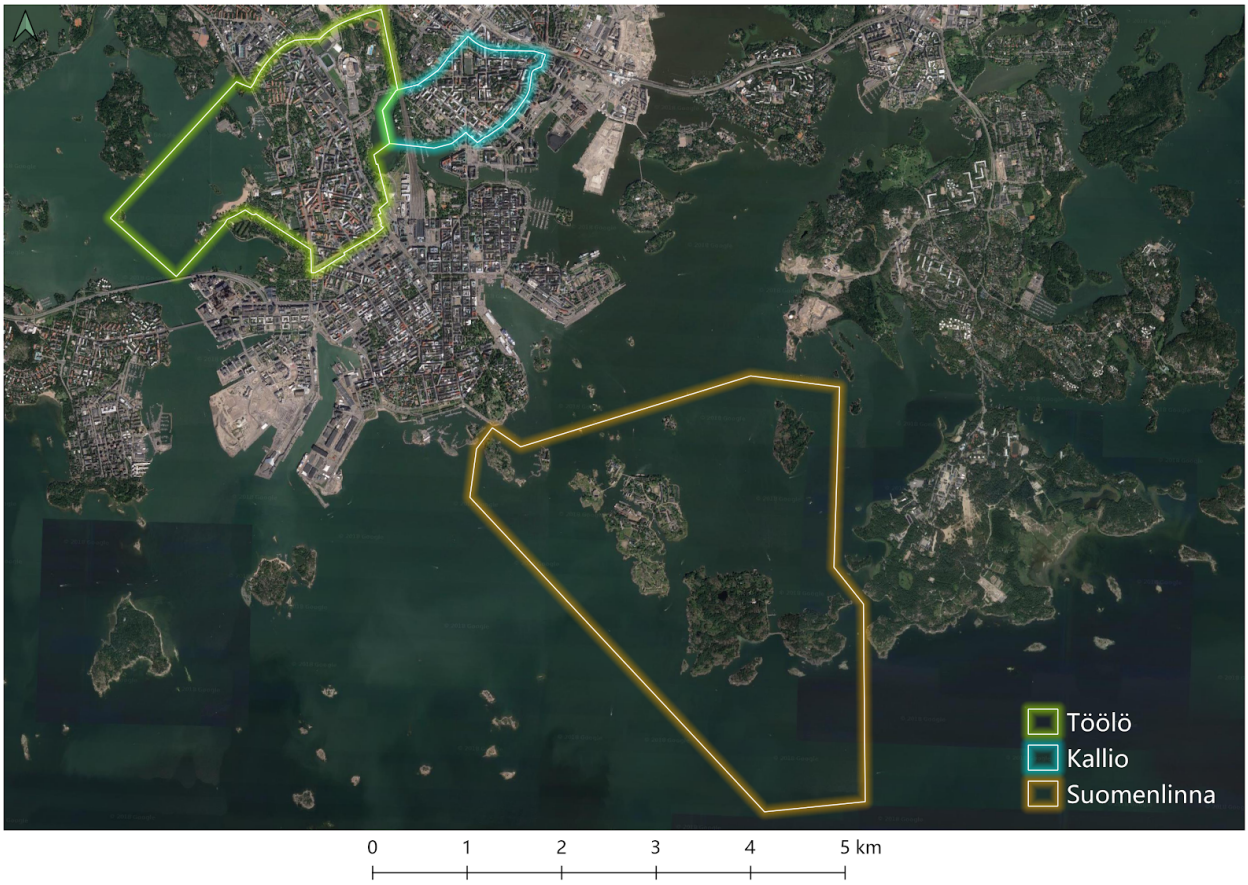
Kuvan 13 kartasta voi nähdä Helsingin katuverkon rakennetta ja johtaa johtopäätöksiä vilkkaimmista alueista Instagram-julkaisujen määrän ja spatiaalisen levinneisyyden avulla, katuverkon rakenne on erityisen hyvin nähtävissä Kampin, Kaartinkaupungin ja Punavuoren välisellä alueella. Kuva 13 on hyvä esimerkki siitä, kuinka Instagram-aineiston kautta näkyvä digitaalinen kaupunkitila heijastaa fyysistä kaupunkitilaa spatiaaliselta rakenteeltaan huolimatta kohdepisteisiin perustuvasta rakenteestaan. Tämä entisestään vahvistaa aineiston sopivuutta digitaalisen kaupunkitilan tutkimiseen. Katuverkon paljastuminen aineistosta on kiintoisaa, sillä se tarkoittaa myös sitä, että kohdepisteet sijoittuvat katujen varsille melko “orgaaniseen” tapaan eivätkä pelkästään rakennusten tai alueiden laskennallisiin keskipisteisiin, joka tekee aineistosta sijainnillisesti luotettavampaa. Pelkästään laskennallisiin keskipisteisiin sidotut sijainnit tekisivät aluekohtaisen tarkastelun tuloksista kappaleessa 3.3.2 hankalasti tulkittavan.



Kuva 13. Tarkempi katsaus raakaan pistemuotoiseen aineistoon Helsingin keskustan alueelta, joka on julkaisumäärältään Helsingin tihein digitaalinen kaupunkitila. Aineistosta erottuu selkeästi fyysisen kaupunkitilan piirteitä, kuten katuverkkoa ja kortteleiden rajoja.

Visuaalisesti tarkasteltuna suurin julkaisukeskittymä kuvassa 13 on Kampin ja Kluuvin välinen alue, joka sisältää muun muassa Helsingin rautatieaseman ja Kampin ostoskeskuksen, jotka ovat erittäin vilkkaita arki- ja vapaa-ajan liikkumisen solmukohtia. Ydinkeskustan ulkopuolelta erottuvia visuaalisia keskittymiä näyttävät olevan Kaapelitehdas Ruoholahdessa, Länsiterminaali Jätkäsaaressa ja Hietalahden torin vieressä sijaitseva Sinebrychoffin puisto. Etu-Töölön alueella pisteet näyttävät olevan varsin tasaisesti levittäytyneitä. Muita tässä mittakaavassa erottuvia keskittymiä on suurien väylien, kuten metroraiteen, varrella. Metroraiteen varrelta korostuvat erityisesti Herttoniemi, Itäkeskus ja Vuosaari. Muita selkeitä keskittymiä ovat

Lauttasaaren itäranta ja Malmi. Arabianrannan varrelle muodostuu selkeä Instagram-julkaisujen sormi. Toinen vastaavanlainen sormi muodostuu Taka-Töölöstä Meilahteen.



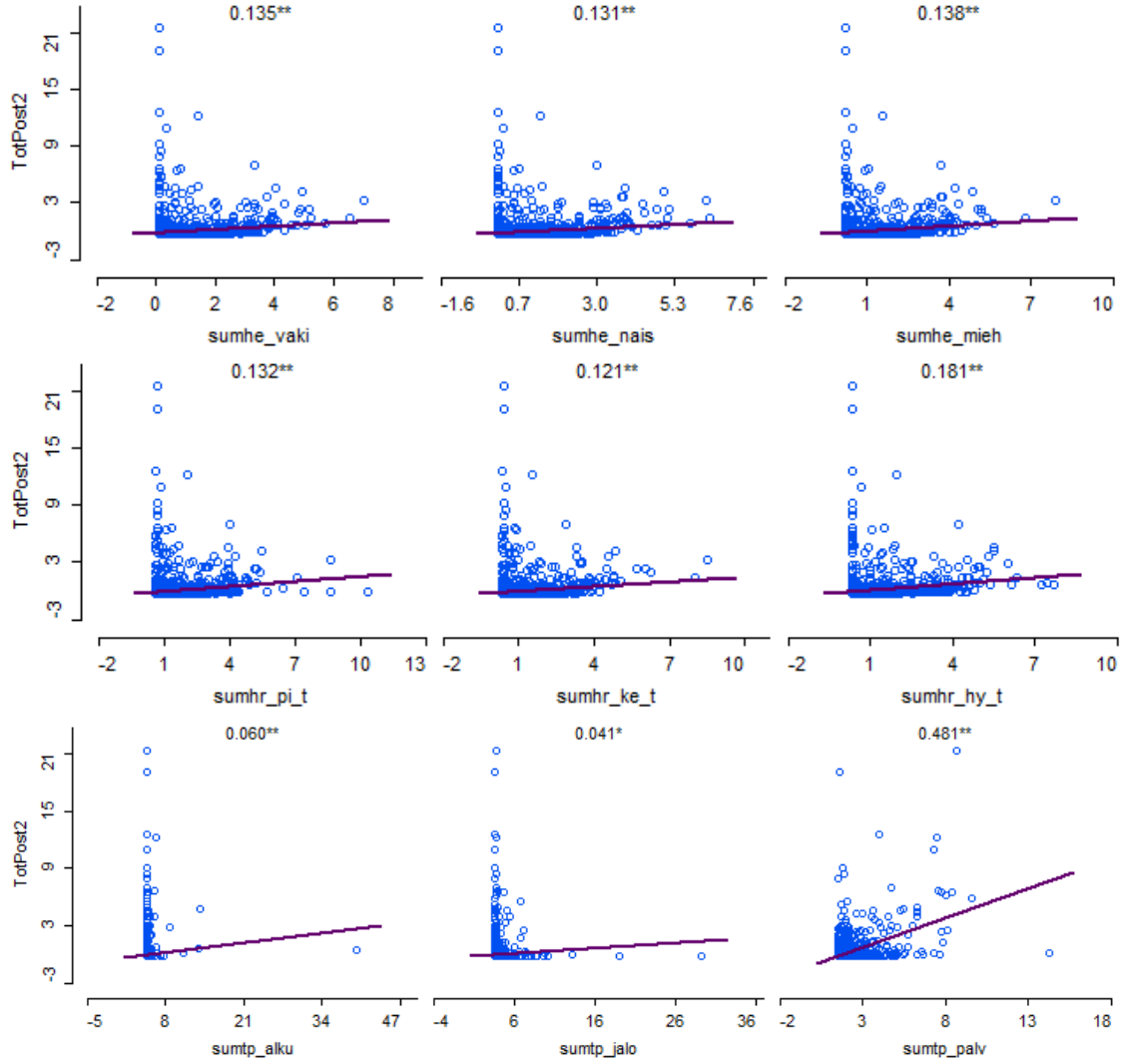
Kuva 14. Lähempään tarkasteluun valitut alueet ovat korostettu tässä kuvassa punaisena hehkuvin reunoin. Käsittelemättömän Instagram-aineiston mukaan Töölössä on noin 100 000 Instagram-julkaisua, Kallion alueella on noin 39 000 julkaisua ja Suomenlinnassa on noin 20 000 julkaisua. Taustan satelliittikuva on Google Earth -palvelusta.

Kuvan 14 kartassa on nähtävissä tässä työssä lähemmässä tarkastelussa olevat kohdealueet: Töölö, Kallio ja Suomenlinna. Kohdealuepohjainen tarkastelu tuonee kieliryhmien väliseen aihevertailuun terävyyttä, joka ei koko Helsingin mittakaavassa tapahtuvassa yleispiirteisemmässä tarkastelussa pääse näkyviin. Töölö valittiin kohdealueeksi sen perusteella, että se on pitkälti asuinalue, jossa sijaitsee myös Helsingin tärkeimpiä nähtävyyksiä, jonka myötä kieliryhmien välisiä eroja voisi olettaa syntyvän. Kallion valitsemisen taustalla on samankaltainen syy, mutta alueelta puuttuu selkeät nähtävyydet ja voisi olettaa alueen olevan hieman enemmän suomea julkaisuissaan käyttävien suosiossa. Suomenlinna on taas alue, joka on harvalle arkinen asuinympäristö, joten kieliryhmien väliset erot eivät liene täällä yhtä selkeitä kuin Töölössä tai Kampissa. Taulukkoon 3 koostetut julkaisumäärät suomeksi ja englanniksi tehdyistä julkaisuista tukevat ennakkokäsitystä kieliryhmien jakautuneisuudesta. Määrät ovat esikäsittelyn ja kielentunnistuksen jälkeen

jäljelle jääneet julkaisut, eivät käsittelemättömät julkaisut. Englanti on julkaisumäärällisesti tarkasteltuna suurempi kieli Töölössä ja Suomenlinnassa, mutta Kalliossa suomi on määrällisesti vahvempi. Tämä voi viestiä siitä, että Kallio on vahvemmin suomenkielellä julkaisevien suosiossa eikä siten näy yhtä vahvasti kansainvälisemmällä kielellä. Tosin on epävarmaa, johtuuko suomenkielisten julkaisujen suurempi määrä Kalliossa mainitusta selkeiden turistinähtävyyksien puutteesta.

Taulukko 3. Valikoitujen kaupunginosien Instagram-julkaisujen määrä kieliryhmittäin.

	suomi	**englanti**
Töölö	18292	20738
Kallio	10563	7310
Suomenlinna	2650	3784



Kuva 15. Kolme standardoitua scatterplot-matriisia julkaisujen lukumäärän sekä YKR-ruudukon (SYKE 2016) asukasmäärän, tuloluokkien ja työpaikkatyyppien mukaan. Ylärivi: väki, naiset ja miehet yhteensä. Keskirivi: pieni-, keski- ja hyvätuloiset yhteensä. Alarivi: Työpaikat alkutuotannossa, jalostuksessa ja palvelusektorilla. Kuvaajien yläpuoliset numerot ovat Pearsonin korrelointiarvoja. Tilastollinen merkitsevyys on ilmoitettu tähdillä: yksi * ($p < 0.05$) sekä kaksi ** ($p < 0.01$).

Kuvassa 15 tarkastellaan julkaisumäärien korrelointia alueellisiin väestömuuttujiin 250 metrin YKR-ruudukossa. Korrelaatio on tilastollisesti merkittävää kaikkien muuttujien osalta, mutta positiivisesti keskivahvaa ainoastaan palvelusektorin työpaikkojen kohdalla. Tämä viestii Instagram-julkaisujen olevan vahvemmin kytketty alueisiin, joilla sijaitsee palveluja, kuin varsinaisiin asuin- tai teollisuusalueisiin. Toisin sanoen, Instagram-julkaisut tapahtuvat palveluihin painottuvilla alueilla kuten liikenteen solmukohdissa ja liikekeskustoissa, sekä ovat lienevät kytköksissä johonkin vapaa-ajan aktiviteettiin, esimerkiksi ruokailuun ravintolassa.

2.2 Aineiston käsittely

Jotta Instagram-aineistoa voi hyödyntää aihemallinnuksessa, julkaisujen kuvatekstit täytyy esikäsitellä huolellisesti. Esikäsittelyssä aineisto ja erityisesti kuvatekstit päätyvät sellaiseen muotoon, jossa tekstien monimuotoisuutta on yksinkertaistettu. Yksinkertaistetut tekstit soveltuvat paremmin kielentunnistukseen, sillä niissä on silloin vähemmän tunnistuksen todennäköisyyttä heikentäviä elementtejä. Kielentunnistuksen jälkeen aineistosta valikoidaan englanniksi ja suomeksi kirjoitetut julkaisut omiksi ala-aineistoikseen ja niiden kuvatekstit lemmatisoidaan. Lemmatisoinnissa tekstejä yksinkertaistetaan edelleen, jotta ne soveltuisivat vektorisoitaviksi, jonka jälkeen LDA-aihemallinnus tulee mahdolliseksi. Aineistoille ajetaan LDA-aihemallinnus eri aluetasoilla, sekä pelkästään substantiiveja ja verbejä sekä pelkästään adjektiiveja hyödyntäen. Aihemalleja syntyy siis yhteensä kaksikymmentä kappaletta eli kymmenen per kieli.

2.2.1 Esikäsittely

Instagram-julkaisujen kuvatekstien esikäsittely on erittäin tärkeä ja aikaa vievä vaihe tässä työssä. Esikäsittelyä on tämän työn vaiheissa kahdessa vaiheessa: ennen kielentunnistusta ja ennen aihemallinnusta. Kuvatekstit ovat sellaisenaan erittäin monimuotoisia käytettyine kielineen, sanoineen, emojineen, hymiöineen ja aihetunnisteineen. Esimerkiksi osa kuvateksteistä on pelkästään emojeita sekä aihetunnisteita. Näiden lisäksi välimerkkien käyttö on vapaamuotoista ja aineistossa onkin esimerkiksi useita peräkkäisiä huutomerkkejä ja kysymysmerkkejä. Kuvateksti voi myös olla kirjoitettu samaan julkaisuun usealla eri kielellä esimerkiksi “Hyvää joulua! Merry Christmas! Frohe Weihnachten!”. Näiden piirteiden

vuoksi kuvatekstejä täytyy yhdenmukaistaa esikäsittelyllä, jotta ensiksi automaattinen kielentunnistus ja toiseksi aihemallinnus onnistuisivat mahdollisimman hyvin (Jauhiainen et al. 2018).

Ennen kuvatekstien käsittelyä aineistosta voi kuitenkin suodattaa yksinkertaisella menetelmällä julkaisuja, jotka todennäköisesti ovat mainoksia tai kilpailuja, joita useat yritykset järjestävät sosiaalisessa mediassa. Työn tarkoitus on selvittää digitaalisen kaupunkitilan eroja suomen- ja englanninkielisissä julkaisuissa Instagramissa, joten edellä mainittujen käyttäjien suodattaminen pois aineistosta tässä vaiheessa on perusteltua. Edellä mainittujen tahojen julkaisut rajattiin aineistosta hyödyntämällä käyttäjänimissä esiintyviä sanoja kuten kauppa, store ja shop (taulukko 4). Tällä mekanistisella suodatuksella aineistosta poistui hieman yli 25 000 julkaisua.

Taulukko 4. Suodatukseen käytetyt sanat ja julkaisumäärät.

Helsinki	Finland	Suomi	.fi	event	shop	store	kauppa	radio	scandinavia	forum
13649	4812	2184	2582	118	859	436	569	589	41	144

Yllä mainitun suodatuksen jälkeen esikäsittelyssä keskitytään seuraavaksi itse kuvateksteihin ja niiden esikäsittelemiseen. Kuten ylempänä mainittiin, jotta kielentunnistus onnistuisi mahdollisimman hyvin, kuvatekstejä tulee yhdenmukaistaa ja niistä poistetaan tunnistusta ja aihemallinnusta haittaavat osat (Jauhiainen et al. 2018), kuten hymiöt, emojit, liialliset välimerkit, maininnat ja aihetunnisteet (taulukko 5). Taulukon 5 kuvatekstit perustuvat oikeisiin kuvateksteihin, mutta ne ovat anonymisoitu Markhamin (2012) ohjeiden mukaisesti siten, ettei niistä pysty tunnistamaan alkuperäisiä käyttäjiä, mutta ne kuitenkin kantavat mukanaan “heijastuksen” oikean kuvatekstin luonteesta. Aihetunnisteet voisivat sisältää paljon aiheisiin liittyvää hyödyllistä tietoa, mutta niiden sisällytys on ongelmallista kirjoitusasunsa vuoksi. Mikäli aihetunnisteet olisivat vain yksittäisiä sanoja, niiden sisällyttäminen olisi mutkatonta, mutta ongelmaksi muodostuu erittäin pitkät, useista sanoista muodostuvat aihetunnisteet, erilaiset kirjoitusasut, kielen vaihtuminen ja kirjoitusvirheet. Yksittäisten sanojen irrottamiseen pitkistä aihetunnisteista eli aihetunnisteiden segmentointiin hyvällä onnistumistarkkuudella on melko monimutkainen prosessi ja siihen

on kehitetty menetelmiä englannille (Srinivasan et al. 2012; Celebi & Ozgur 2017), mutta suomelle ei. Näistä syistä aihetunnisteiden segmentointia ei hyödynnetä tässä työssä.

Mikäli kuvatekstejä käytettäisiin myös sävyanalyysiin, emojien ja hymiöiden sisällyttäminen olisi erityisen arvokasta, sillä ne sisältävät paljon tekstin ja julkaisun sävyyn liittyvää tietoa. Sävyanalyysiä ei kuitenkaan tässä työssä tehdä, sillä suomenkielisiä sävymalleja ei ole avoimesti saatavilla ja sellaisen luominen työtä varten on tämän työn mittakaavan ylittävä pyrkimys. Lisäksi sävymallin täytyy sopia käyttötarkoitukseen, esimerkiksi tuote- tai elokuva-arvostelujen perusteella luotu sävymalli ei välttämättä sovellu kovinkaan hyvin sosiaalisen median julkaisujen sävyjen ja subjektiivisuuden analysoimiseen (Taboada et al. 2011). Kuvatekstien yhdenmukaistamisen jälkeen julkaisujen kuvateksteissä käytetyt kielet tunnistetaan, niistä valitaan englanniksi ja suomeksi kirjoitetut kuvatekstit, jonka jälkeen valitut kuvatekstit lemmatisoidaan ja lopulta lemmatisoitujen kuvatekstien perusteella luodaan LDA-aihemallit.

Taulukko 5. Esimerkkejä kuvitteellisista kuvateksteistä muokkaamattomina ja esikäsitellyssä asussa, jossa ne syötetään automaattiseen kielentunnistukseen. Esikäsitellyistä kuvateksteistä on poistettu aihetunnisteet, maininnat, emojit ja hymiöt.

Kuvateksti	Esikäsitelty kuvateksti
Uusi vuosi, uudet verkkarit ja perhe! #uusivuosi #2016 #newyear #love #family #fun #perhe #rakkaus #baby	uusi vuosi, uudet verkkarit ja perhe!
... ja vuoden eka lenkki 💪 #malmi #airport	ja vuoden eka lenkki
#Lokki syö hanhenmaksaa ja huutelee tarjoilijalle	syö hanhenmaksaa ja huutelee tarjoilijalle
Thanks guys we had a great time! @sportsteam #nyrillataan	thanks guys we had a great time!
Fireworks.. It's selfie time! 😂 #winter2015 #Helsinki #newyear2016 #NewYearsEve #party #selfie #socialmedia #time #brunettes #girls	fireworks it’s selfie time
fanatic = someone who will not change his mind or the subject #mykindofbeer	fanatic someone who will not change his mind or the subject

2.2.1.1 Automaattinen kielentunnistus

Menetelmän kuvaus

Automaattinen kielentunnistus on kieliteknologinen menetelmä, joka pyrkii tunnistamaan tekstissä käytetyn kielen (Jauhiainen et al. 2018). Tähän työhön käytettävissä olevia kielentunnistuskirjastoja oli kolme: fastText, CLD2 ja langid. Näiden tarkkuudessa ja nopeudessa on selkeitä eroja. Facebookin kehittämä

fastText on osoittautunut parhaimmaksi tarkkuudeltaan ja nopeudeltaan tämän työn kaltaisessa tutkimuksessa (Hiippala et al. 2018), joten se valikoitui siten myös tähän työhön. Edellä mainitun kaltaisesti puhdistetut tekstit (taulukko 5) syötetään fastText-kirjastolle, joka luokittelee tekstit kuvatekstikohtaisesti jollekin kielelle. Kieliluokittelun lisäksi luokitukseen tallentuu myös tunnistuksen todennäköisyys ja tunnistamiseen käytetyn tekstisyötteen pituus, johtuen tässä työssä käytetystä Hiippalan ja muiden (2018) kehittämästä fastText-kirjastoa hyödyntävästä työkalusta. Näin ollen yhdestä kuvatekstistä voidaan tunnistaa useampia kieliä, mikäli eri virkkeet on kirjoitettu eri kielillä. Saman virkkeen sisäinen kielen vaihtelu pudottaa tunnistamisen varmuutta, eikä sitä siis tunnisteta kahdeksi eri kieleksi. Esimerkiksi teksti “After watching Napapiirin Sankarit film we went for a few delicious beers, kippis!” luokittuisi englanniksi mutta heikommalla todennäköisyydellä kuin täysin englanniksi kirjoitettu kuvateksti. Kun taas teksti “Hieno auringonlasku merellä! Nice sunset at sea!” luokittuisi suomeksi ja englanniksi.

Automaattinen kielentunnistus ei ole täysin virheetön menetelmä ja sosiaalisen median julkaisujen kirjavan slangin-, kielenkäytön ja kielten vaihtelun myötä tunnistusvirheiden mahdollisuus lienee selkeästi suurempi kuin esimerkiksi viranomaistekstin kanssa (Carter et al. 2013, Jauhiainen et al. 2018). Näin ollen aihemallin tulosten parantamiseksi luotiin suodatin kriteerit julkaisujen valintaan: kielentunnistuksen jälkeen vain ne suomen- ja englanninkieliset julkaisut valikoituivat, joiden kielen tunnistuksen todennäköisyys oli yli 70 % ja syötteen pituus enemmän tai yhtä suuri kuin 8 merkkiä. 70 % varmuus ei ole liian tiukka, mutta joka kuitenkin selkeästi hylkää sellaiset julkaisut, jotka ovat liian epävarmasti tunnistettuja sekä ne, jotka ovat kirjoitettu puoliksi yhdellä kielellä ja puoliksi toisella kielellä. Käytetty 8 merkin raja perustuu Hiippalan ja muiden (2018) hyväksi toteamaan rajaan heidän vertaillessa eri automaattisia kielentunnistusmenetelmiä, vaikkakin tämä raja ei todennäköisesti ole yleistettävissä koskemaan automaattista kielentunnistusta yleispiirteisesti (Jauhiainen et al. 2018). Tällä tavalla minimoidaan lyhyiden kuvatekstien mukana tulevan virheen vaikutus, sillä lyhyillä kuvateksteillä automaattiset kielentunnistusmenetelmät ovat varsin epävarmoja. Esimerkiksi useassa kielessä samassa kirjoitusasussa olevat sanat vaikeuttavat tunnistusta automaattista kielentunnistusta käytettäessä. Lyhyiden kuvatekstien kuten “Noni”, “Hei” ja “Auto” kieli voidaan tunnistaa väärin niin automaattisesti, mutta myös manuaalisesti (ihmisvoimin) annotoidessa.

Esimerkiksi Auto-sana esiintyy sellaisenaan lukuisissa eri kielissä, joten pelkästään sen perusteella on mahdotonta sanoa mitä kieltä kuvatekstissä on käytetty. Samankaltainen luokittelun kannalta epäselvä tilanne tulee vastaan myös paikannimien, esimerkiksi sanan “Helsinki”, kanssa kuten Hiippala ja muut (2018) ovat todenneet. Tämänkaltaisten virheiden minimoimiseksi kyseinen suodatin otettiin käyttöön aineiston valinnassa.

Kielentunnistus Instagram-aineistolle

Automaattisessa kielentunnistuksessa kieliä tunnistettiin onnistuneesti 659 323 julkaisun osalta, kun kokonaismäärä julkaisuja on 828 588 kappaletta. Julkaisuista, joista tunnistettiin kieli (659 323), vain yhtä kieltä käytettiin 93,9 % julkaisuista (619 078) ja kahta tai useampaa kieltä vain 6,1 % julkaisuista. Kielen vaihto tunnetaan kielitieteessä koodinvaihtona (engl. *code switching*) ja julkaisujen sisällä se ei vaikuta olevan kovin yleistä tässä Instagram-aineistossa. Tosin kielten välillä vaihtelu tallentuu tässä menetelmässä ainoastaan silloin, jos eri kielillä kirjoitetut virkkeet ovat erotettu toisistaan välimerkein. Mikäli yhden virkkeen sisällä käytetään useampaa kieltä, kielen tunnistamisen todennäköisyys alentuu ja vieraskielinen sana jää lauserakenteeseen. Yksittäisten vieraskielisten sanojen siivoaminen aineistosta on käytännössä mahdotonta, sillä se vaatisi kattavan ja kontekstiriippuvaisen sanalistan, jonka muodostaminen on käytännössä mahdotonta. Näin ollen esimerkiksi lähes täysin suomeksi kirjoitetussa kuvatekstissä olevat muutamat englantilaiset sanat tiputtavat menetelmän luokituksen varmuutta, prediktiota, vain hieman. 660 000 kielitunnistetussa julkaisussa on muutamia yllä mainittuja epävarmuustekijöitä mukana, joiden vaikutus pyrittiin minimoimaan työssä merkki- ja tunnistusvarmuusrajalla.

Lemmatisointia varten kielitunnistetusta aineistosta valittiin ne julkaisut, joiden tunnistukseen käytetyn tekstin pituus oli 8 tai enemmän merkkiä, koska lyhyiden tekstien kielentunnistamisen varmuus on varsin heikkoa (Jauhiainen et al. 2018). Tämän myötä kielitunnistetusta aineistosta valikoitui 87 % eli 574 482 kappaletta. Kun näistä valikoiduista suodatettiin pois alle 70 % varmuudella tunnistetut julkaisut, kielitunnistetusta aineistosta on jäljellä 61 % eli 403 095 julkaisua. Toisin sanoen, alkuperäisestä noin 828 000 julkaisun aineistosta yli puolet rajautuu pois tämän tutkielman kannalta käyttökelvottomina esikäsittelemisen ensimmäisessä vaiheessa. Aineiston puolittuminen on harmillista, sillä analyysin kattavuus

kärsii, mutta asetettujen laatukriteerien läpäissyt reilun 400 000 julkaisun aineisto on silti huomattavan suuri. Tästä suodatetusta aineistosta valittiin suomen- ja englanninkieliset julkaisut omiksi erillisiksi aineistoikseen kuvatekstien lemmatisointia varten. Suodatuksen jälkeen kävi ilmi, että jäljelle jäänyt suomenkielinen aineisto on määrällisesti suurempi kuin jäljelle jäänyt englanninkielinen aineisto: suomenkielinen aineisto sisältää noin 180 000 ja englanninkielinen 162 000 julkaisua. Suomeksi tehtyjen julkaisujen nousun taustalla lienee virkkeen sisäistä koodinvaihtoa sisältäneet kuvatekstit. Tämä kertoo mahdollisesti siitä, että pääasiallisesti suomeksi kirjoitettuihin kuvateksteihin ei sekoiteta muita kieliä yhtä voimakkaasti kuin pääasiallisesti englanniksi kirjoitettuihin kuvateksteihin. Aineiston vaiheittainen pieneneminen esikäsittelemisen ja suodatuksen myötä on koostettu taulukkoon 6. Esikäsittelyssä kuvatekstien pituus lyhentyi noin keskimäärin noin 50 % molemmilla kielillä aihetunnisteiden, mainintojen, emojien ja muiden elementtien poistamisen myötä.

Taulukko 6. Taulukkohavainnollistus aineiston vaiheittaista rajautumista esikäsittelyssä ja automaattisen kielentunnistuksen jälkeen. Ylempi luku kuvaa julkaisujen määrää ja alempi luku osuutta alkuperäisestä aineistosta.

Alkup. aineisto	**Tunnistettu kieli**	**Yksi kieli**	**Suomi**	**Englanti**	**Suomi (p >= 0.7)**	**Englanti (p >= 0.7)**
828 588	659 323	619 078	219 797	304 690	180 032	162 189
100%	79.57%	74.71%	26.53%	36.77%	21.73%	19.57%

Automaattisessa kielentunnistuksessa käytettyyn menetelmään voisi saada lisää tarkkuutta, mikäli siinä pystyisi huomioimaan saman käyttäjän muut julkaisut, jonka myötä ristiriitatilanteissa aikaisemmin käytetyt kielet saisivat suuremman painoarvon (Carter et al. 2013). Esimerkiksi jos käyttäjä A on kaikissa muissa julkaisuissaan käyttänyt joko suomea tai englantia kielenään, niin hän tuskin käyttää näistä poikkeavaa kieltä, kuten baskia tai arameaa yhdessä virkkeessään. Tällaisen kontekstia hyödyntävän menetelmän puuttuessa väärin tunnistettuja virkkeitä jää auttamatta aineiston ulkopuolelle sekä sisältyy aineistoon. Tosin julkaisuja, joissa on fastTextin mukaan käytetty useampaa kuin yhtä kieltä, on melko vähän: vain 40 245 kappaletta. Näistä reilusta 40 000 vain hieman alle 10 000 on yli 70 % varmuudella tunnistettuja eli kovin suuresta virhe-elementistä ei varmaan ole kyse. Näissä julkaisuissa on käytetty yhtä tai useampaa fastTextin tunnistamaa kieltä.

2.2.1.2 Kuvatekstien lemmatisointi

Menetelmän kuvaus

Lemmatisoinnissa sanan taipunut muoto muutetaan sen perusmuotoon. Tämä on havainnollistettu esimerkein taulukossa 7. Lemmatisointi on tärkeä edellytys aihemallinnuksen tuloksien onnistumisen kannalta, joka korostuu morfologisesti rikkaiden kielien (May et al. 2016), kuten suomen, kanssa. Morfologisesti rikas kieli tarkoittaa kieltä, jossa sanoilla on useita taipuneita muotoja. Taipuneet sanat tekevät suomenkielisestä tekstistä hankalan tekstin aihemallinnuksen kannalta, sillä aihemallinnus ei automaattisesti ymmärrä sanojen taipuneita muotoja samaksi sanaksi, vaan käsittelee ne erikseen sillä oletuksella, että ne tarkoittavat eri asioita, vaikka todellisuudessa näin ei ole. Tämä tuo aineistoon suuren määrän "hälinää", kuten samojen sanojen taipuneita muotoja ajassa ja eri persoonissa, ja tekee sellaisen aineiston aihemallinnuksen tuloksen käytännössä käyttökelvottomaksi. Esimerkiksi koti-sanan taipuneita muotoja ("kotiin", "kodista", "kotimme") käsiteltäisiin täysin eri sanoina, joita aihemallinnus ei osaa luokitella samaksi sanaksi. Lemmatisoimattomien tekstien kanssa on lähes varmaa, että suuri osa aiheista, jotka muuten lemmatisoidun tekstin aihemallissa nousisivat selkeästi esille, jäävät näkymättömiin kielen morfologisen rikkauden vuoksi. Tämän vuoksi erityisesti suomenkielisen tekstin lemmatisoiminen on erityisen tärkeää aihemallinnuksen kannalta.

Englannin osalta lemmatisointi ei ole yhtä välttämätön toimenpide, koska morfologinen rikkaus englannissa on varsin vähäistä esimerkiksi suomeen tai venäjään verrattuna (May et al. 2016). Tästä piirteestä huolimatta on erittäin suositeltavaa, että myös englanninkieliset tekstit lemmatisoidaan, sillä se yhdenmukaistaa aineistoa sisäisesti ja tekee tuloksista paremmin vertailukelpoiset suomenkielisen aineiston kanssa.

Taulukko 7. Esimerkkejä lemmatisoinnin lopputuloksista erilaisille kuvitteellisille kuvateksteille suomeksi ja englanniksi.

Kuvateksti	Esikäsitelty + Lemmatisoitu
Uusi vuosi, uudet verkkarit ja perhe! #uusivuosi #2016 #newyear #love #family #fun #perhe #rakkaus #baby	uusi vuosi uusi verkkari ja perhe
... ja vuoden eka lenkki💪 #malmi #airport	ja vuosi eka lenkki
#Lokki syö hanhenmaksaa ja huutelee tarjoilijalle	syödä hanhenmaksa ja huudella tarjoilija
Thanks guys we had a great time! @sportsteam #nyrillataan	thank guy we have a great time
Fireworks.. It's selfie time! 😁 #winter2015 #Helsinki #newyear2016 #NewYearsEve #party #selfie #socialmedia #time #brunettes #girls	firework it be selfie time
fanatic = someone who will not change his mind or the subject #mykindofbeer	fanatic someone who will no change he mind or the subject

Hyvin suunniteltu lemmatisointimenetelmä tai -työkalu ottaa lemmatisoinnissa huomioon tekstissä esiintyvien sanojen kontekstin. Konteksti muodostetaan lemmatisoinnin kanssa samanaikaisesti tapahtuvasta sanojen sanaluokkajäsennyksestä (engl. *part-of-speech tagging*), joka huomioi tietyn määrän lemmatisoitavaa sanaa edeltäviä ja seuraavia sanoja, jotta mahdolliselta virheelliseltä lemmatisoinnilta vältytään. Tämä on hyödyllistä esimerkiksi silloin kun useampi eri sana voi olla ulkoasultaan identtinen. Yleinen esimerkki moniselitteisestä suomenkielisestä sanasta on kuusi, joka voi tarkoittaa numeroa, puuta ja Kuuta (2. persoonan omistusliitteellä). Lisäksi sana voi myös olla polyseeminen, jolloin se on merkitykseltään läheinen, mutta tarkoittaa käyttöyhteydessä eri asioita (Tieteen termipankki 2018), kuten mennä-verbin kanssa: “mennä hyvin” ja “mennä kotiin”. Sama ilmiö esiintyy englannin kielessä. Esimerkiksi meeting-sanan kanssa, joka voi tarkoittaa sekä kokous-substantiivia (*a meeting*) tai tavata-verbin infinitiiviä (*to meet*). Sanaluokan jäsennystä voi myös hyödyntää siten, että suuresta määrästä tekstejä voi irrottaa esimerkiksi kaikki verbit tai adjektiivit, kuten tässä työssä tehdään substantiivi-verbi- ja adjektiivi-aihemallinnuksien kanssa. Näiden seikkojen vuoksi lemmatisointi on mekanistisempia yksinkertaistusmenetelmiä, kuten stemmausta, parempi tapa esikäsitellä tekstiä ennen aihemallinnusta.

Lemmatisointi Instagram-aineistolle

Tässä työssä lemmatisointiin käytettiin kahta työkalua, FinnPOS:ia (Silfverberg et al. 2016) ja spaCy:ä (ExplosionAI 2018), joiden käyttötapa ja -vaikeus olivat toisistaan selkeästi erottuvat. FinnPOS on avoin työkalupaketti suomenkielisten tekstien lemmatisointiin ja sanaluokan leimaukseen. SpaCy on

Python-ohjelmointikielelle suoraan rakennettu luonnollisen kielen prosessoinnin työkalukirjasto, jonka käyttö ja ohjeistus ovat erittäin selkeitä ja helppoja. FinnPOS:n käyttämisen vaikeus johtui pitkälti sen vaatimasta Linux- tai OS X-käyttöjärjestelmästä, jonka myötä työkalun käyttäminen Windows-käyttöjärjestelmässä ei ollut mahdollista. Tämän vuoksi suomenkielisten kuvatekstien lemmatisointi toteutettiin etäyhteydellä CSC:n Taito-superklusterissa.

2.2.2 Aihemallinnus LDA-menetelmällä

Menetelmän kuvaus

Aihemallinnus on kieliteknologinen menetelmä suurten tekstimassojen analyysiin (Blei 2012a; Martin & Schuurman 2017). Eräs yleisimmin käytetyistä aihemallinnusmenetelmistä on vuonna 2003 kehitetty Latent Dirichlet Analysis eli LDA-malli, joka on ohjaamaton todennäköisyyksiin perustuva luokittelumalli (Blei et al. 2003a; Blei 2012a, 2012b; Lansley & Longley 2016). Aihemallinnus voi tuoda huomattavia helpotuksia kvalitatiivisen paikkatiedon analyysin ja visualisointiin, sillä se on huomattavasti nopeampaa verrattuna käsin tehtävään tekstien aiheiden luokitteluun (Blei et al. 2003a; Lau et al. 2011; Blei 2012a; Martin & Schuurman 2017).

LDA-aihemallinnuksessa malli käy sille annetun kokoelman tekstidokumentteja eli korpuksen, tässä tapauksessa julkaisujen kuvatekstit, läpi useaan otteeseen vertaillen dokumenteissa esiintyvää sanastoa ja sanojen käyttöä, sekä mallintaen näiden välisistä suhteista esiin nousevia piileviä eli latentteja aiheita (Blei 2012a, 2012b; Fu et al. 2018). Aiheet muodostuvat dokumenteissa esiintyvien uniikkien sanojen esiintymistiheyksien ja levinneisyyksien perusteella. Esimerkiksi usein samassa dokumentissa esiintyvät sanat alkavat mallin “rakentuessa” muodostamaan aihekokonaisuutta. Näitä aihekokonaisuuksia muodostetaan käyttäjän määrittelemän määrän verran. Kun Instagram-julkaisuja, eli aihemallille annettuja dokumentteja, luokitellaan aiheisiin todennäköisyyksien perusteella, tulee huomioida, että dokumentti voi kuulua useampaan aiheeseen. Esimerkiksi julkaisun kuvateksti voi mallin mukaan kuulua 70% aiheeseen 1 ja 30% aiheeseen 15. Dokumentin kuuluessa useaan aiheeseen sen aiheluokituksen todennäköisyys laskee, kuten taulukon 8 alimmassa kohdassa tapahtuu. Joissain tapauksissa dokumentti ei kuuluu mihinkään mallinnettavaan aiheeseen varsinaisesti, jolloin sen todennäköisyys kuulua kaikkiin mallinnettaviin aiheisiin

on yhtä suuri LDA-mallinnuksen oletusasetuksilla. Tällöin esimerkiksi viiden aiheen aihemallissa kuvatun kaltainen dokumentti saisi kaikkien aiheiden osalta 0.20 todennäköisyyden, eikä siten ole erityisesti minkään aiheen “oma”. Sosiaalisen median julkaisut antavat myös oman haasteensa aihemallinnukselle lyhyen pituutensa kautta (Hong & Davison 2010).

Aihemallinnus vaatii sanojen muuntamista numeeriseen muotoon, jolloin tekstille tehtävät matemaattiset ja tilastolliset toimenpiteet mahdollistuvat. LDA-aihemallinnus käyttää sanaston *bag-of-words* -vektorirepresentaatioita, joissa jokainen uniikki sana saa sitä vastaavan tunnusluvun ja frekvenssiarvon kokonaislukuna, joka kuvaa kuinka yleinen kyseinen sana on koko korpuksessa (Rehurek 2018). LDA-aihemallinnuksen perusoletuksena on, että mallille annettavien dokumenttien tekstit sisältävät useita aiheita ja aiheet muodostuvat dokumentissa usein esiintyvien sanojen levinneisyyksien perusteella (Blei et al. 2003a; Blei 2012a, 2012b). Kuva 16 havainnollistaa mallin toimintaperiaatetta. Lopputuloksena gensim-kirjaston LDA-aihemallinnusmenetelmä tuottaa käyttäjän määrittelemän lukumäärän aiheita sen läpikäymästä korpuksesta sekä jokaiselle aiheelle käyttäjän määrittelemän määrän tärkeimpiä sanoja.

Menetelmä ei nimeä mallintamiaan aiheita termein kuten autoilu, viihde ja ruoka vaan pelkästään numeroin. LDA-aihemallien aiheiden nimeäminen on ollut ongelmallista aihemallinnuksessa (Lau et al. 2011; Blei 2012a; Aletras et al. 2014). Kirjallisuudessa on ollut käytäntönä nimetä aihe subjektiivisesti kymmenen aiheelle tärkeimmän sanan avulla helpottamaan ja havainnollistamaan tuloksia lukijalle (Lau et al. 2011; Aletras et al. 2014), eikä niinkään osoittamaan tieteellisesti, että jokin nimi kuvaa jotain aihetta tilastollisesti merkittävällä tasolla. Tämä heikentää aihemallinnuksen toistettavuutta, mutta havainnollistamisen kannalta nimeäminen on lähes välttämätöntä. LDA-mallista on tosin kehitetty myös muunnoksia, jotka yrittävät luoda aiheille nimiä automaattisesti (Lau et al. 2011; Aletras et al. 2014; Martin & Schuurman 2017) sekä puhtaasti tekstidokumenttien aiheiden nimeämiseen kehitettyjä työkaluja (Maiya et al. 2013), mutta joko ne eivät vaikuta käyttökelpoisilta sosiaalisen median tekstiaineistoille, vaativat jatkokehittämistä tai ne eivät ole vapaasti käytettävissä (Lau et al. 2011; Maiya et al. 2013).

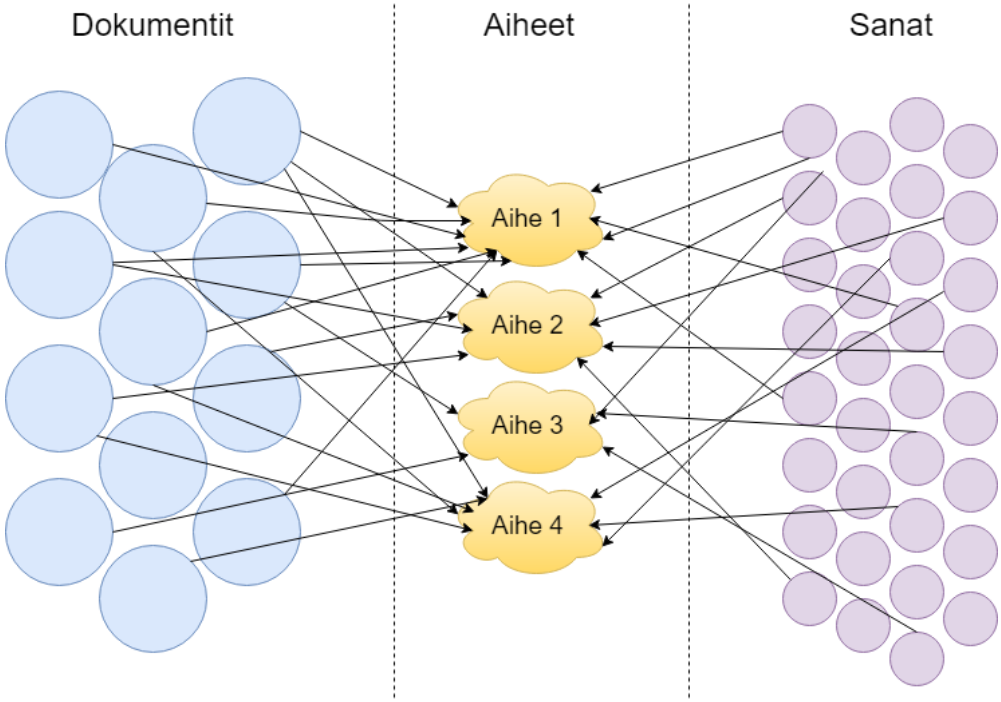
Aiheiden lukumäärän määrittely riippuu aihemallille annettavasta aineistosta (Blei 2012a; Aletras et al. 2014). Esimerkiksi Lansleyn & Longleyn artikkelissa (2016) hyväksi aiheiden lukumääräksi suurkaupungin kokoluokan analyyseissä todettiin 20, kun esikäsittelyn jälkeen geoleimattuja Twitter-julkaisuja oli jäljellä noin 1,3 miljoonaa. Käytössä oleva Instagram-aineisto on sisällöltään melko homogeeninen, lukumäärältään pienempi (suomenkieliset julkaisut noin 180 000 ja englanninkieliset noin 160 000) ja maantieteellisesti pienemmältä alueelta verrattuna Lansleyn ja Longleyn (2016) Twitter-aineistoon, joten pienempi aihemäärä soveltunee tähän työhön ja tälle aineistolle paremmin. Lisäksi suuri mallinnettavien aiheiden määrä vaikeuttaa tuloksien tulkintaa ja havainnollisuutta (Fu et al. 2018), jonka vuoksi mallinnettavien aiheiden määrä pyritään pitämään melko alhaisena. Pienen aihemäärän mukana tulkinnallisuus paranee, mutta aiheiden erityisyys heikkenee jolloin mallintuneet aiheet ovat todennäköisemmin yleispiirteisiä. Esimerkiksi pienellä 10 aiheen aihemäärällä muodostettu aihemalli saattaa muodostaa yksittäisen yleispiirteisen “ruoka”-aiheen, kun taas 300 aiheen aihemallilla “aamiainen”, “ruokakauppa”, “sushi” ja “pizza” voisivat olla omia aiheitaan. Myös spatiaalisen mittakaavan on todettu vaikuttavan aiheiden yleispiirteisyyteen, jolloin samalla aihemäärällä mutta kahdella toisistaan selkeästi eroavalla mittakaavatason tarkastelulla aiheiden yleispiirteisyys nousee suurien aluetarkasteluiden kanssa ja laskee kun tarkastellaan pieniä alueita (Jenkins et al. 2016).

Taulukko 8. Latent Dirichlet Allocation -menetelmän toimintaperiaatteesta yleisesti käytetty havainnollistus, jossa viidestä dokumentista mallinnetaan kaksi aihetta (Chen 2011).

Lauseet	**Aihe 1**	**Aihe 2**
Tykkään syödä parsakaalia ja banaania	100 %	0 %
Söin banaanin ja smoothien aamiaiseksi	100 %	0 %
Chinchillat ja kissat ovat söpöjä	0 %	100 %
Siskoni adoptoi kissanpennun	0 %	100 %
Söpö hamsteri syö parsakaalia	60 %	40 %

Kaikkia kuvatekstien sanoja hyödynnetään kaikissa aihemallinnuksissa, mutta aihemallinnus ajetaan erikseen myös pelkästään kuvatekstien adjektiiveille sekä kuvatekstien verbeille ja substantiiveille kaupunginosatarkastelussa. Adjektiivien aihemallinnuksella kokeillaan erittäin kevyen sävyanalyysin toteuttamista, jonka perusteella mahdollisesti saadaan käsitys siitä, miten eri sävyillä tehdyt Instagram-julkaisut jakautuvat kohdealueilla. Tosin adjektiivien vahva painottuminen positiivissävytteisyyteen todennäköisesti tekee tuloksista yllätyksettömiä ja syntyvien “sävyaiheiden”

semanttisesta erottelusta ja mahdollisesta pisteytyksestä mahdotonta. Verbien ja substantiivien mallinnuksella pyritään saamaan selkeyttä aiherakenteeseen, jota adjektiivien sisällytys voi sekoittaa, sillä adjektiivit liittyvät vahvemmin sävyyn kuin aihesisältöön, mutta myös tuomaan julkaisujen aktiviteetit paremmin havaittaviksi. Näin parhaimmassa tapauksessa muodostuva “kolmitahoinen” (aiheet, sävyt ja aktiviteetit) kuva digitaalisesta kaupunkitilasta saa hieman syvyyttä, jota ei voisi saavuttaa pelkästään yhdellä aihemallilla. Aiheita, sävyjä ja aktiviteetteja voisi käyttää Jan Gehlin (2011) luoman laatukehikon (taulukko 1) kautta ympäristön laadun arvioimiseen digitaalisen kaupunkitilan kautta. Näin ollen alueet, joilla valinnaiset ja sosiaaliset aiheet ovat selvässä enemmistössä, voivat merkitä laadukasta kaupunkiympäristöä, joka havaitaan digitaalisen kaupunkitilan kautta. LDA-analyysi toteutetaan koko Helsingin kattavalle aineistolle hyödyntäen julkaisujen kuvatekstejä kokonaisina sekä rajatuille kolmelle kohdealueelle, joihin sovelletaan myös mainittua adjektiivi- sekä verbi/substantiivimallinnusta. Kohdealueiden aihemallinnuksella pyritään saada spatiaalisesti rajattuja ja aiheellisesti keskittyneempiä aihekokonaisuuksia kuin koko Helsingin aihemallinnuksella, jonka voisi olettaa olevan melko yleispiirteinen lopputulokseltaan aikaisempien tutkimuksien valossa (Jenkins et al. 2016).



Kuva 16. Havainnollistus LDA-aihemallinnuksen toimintaperiaatteesta. Dokumentteja käsitellään kokoelmana sanoja ja samanaikaisesti kaikkien dokumenttien sanoja käsitellään dokumenteista irrallisina objekteina. Aiheet muodostuvat dokumenttien ja sanojen tilastollisista suhteista toisiinsa. Sama sana ja dokumentti voi kontribuoida yhteen tai useampaan aiheeseen saman mallin sisällä, kuten taulukosta 8 selviää.

Aihemallinnus Instagram-aineistolle

Kappaleessa 2.2.1 esitellyn esikäsittelemisen jälkeen molempien aineistojen kuvatekstien sanat tokenisoidaan, joka tarkoittaa sanojen käsittelyä erillisinä objekteina eikä yhtenä merkkijonona tai kokoelmana kirjaimia. Tokenisoinnin jälkeen tokenisoidut sanat muunnettiin bag-of-words -vektorin mukaiseen numeeriseen muotoon LDA-aihemallia varten. Muodostetuista sanastoista poistettiin kaikki sanat, jotka esiintyvät vain kerran koko aineistossa ja vain 1-2 kirjaimesta koostuvat sanat. Lisäksi sanoista hylättiin ne sanat, jotka esiintyvät vain 2 julkaisussa sekä sanat jotka esiintyvät yli 50 % kaikista kuvateksteistä. Suodatus pyrkii parantamaan aihemallin tarkkuutta: harvoin esiintyvät sanat eivät heikennä muodostuvien aiheiden vahvuuksia, eivätkä usein esiintyvät sanat dominoi aiheita ja siten peitä muita aiheita näkymättömiin.

Suomenkielisissä julkaisuissa (180 032 kappaletta) on yhteensä 925 791 kappaletta sanoja, jota koostuvat 36 733 uniikista sanasta. Toisin sanoen suomenkielisessä aineistossa uniikit sanat esiintyivät keskimäärin 25 kertaa. Englanninkielisten julkaisuissa (162 189 kappaletta) on yhteensä 741 286 sanaa, jotka koostuvat 21 197 uniikista sanasta eli uniikit englanninkieliset sanat esiintyivät aineistossa keskimäärin 35 kertaa. Tässä yhteydessä uniikki sana ei tarkoita vain kerran aineistossa esiintyvää sanaa vaan niitä sanoja, joista julkaisujen kuvatekstit muodostuvat. Esimerkiksi sana ‘auto’ on uniikki sana, joka voi esiintyä aineistossa noin 5 000 kertaa useassa julkaisussa, kun taas sana ‘automekaanikko’ on uniikki sana, joka voi esiintyä aineistossa vain kerran yhdessä julkaisussa, jonka myötä sitä ei sisällytetä aihemalliin yllä kuvatun suodatuksen vuoksi. Aihemallinnuksien tulokset esitetään taulukoin, kuvaajin ja kartoin.

Taulukko 9. Koko Helsingin kattavalle suomenkieliselle aineistolle luodut LDA-aihemallit, joiden perusteella aineistolle sopivimmat parametrit valittiin. Mukana on myös samalta alueelta toteutettu pelkästään substantiiveihin ja verbeihin keskittyvä aihemalli, jossa käytetään iteraatio- ja läpikäyntivertailuissa hyviksi todettuja parametreja.

Iteraatiovertailu				
Aiheita	10	10	**10**	10
Iteraatioita	200	900	**1400**	10000
Läpikäyntejä	5	5	**5**	5
Koherenssi (C_V)	0.25	0.28	**0.29**	0.29
Läpikäyntivertailu				
Aiheita	10	**10**	10	10
Iteraatioita	1400	**1400**	1400	1400
Läpikäyntejä	20	**30**	50	90
Koherenssi (C_V)	0.26	**0.31**	0.28	0.29
	Subs-verb.	Adjektiivit	Subs-verb.	Adjektiivit
Aiheita	10	10	10	10
Iteraatioita	1400	1400	1400	1400
Läpikäyntejä	30	30	60	60
Koherenssi (C_V)	0.26	0.39	0.25	0.59

Koko Helsingin alueelle luotaville suomen- ja englanninkielisille LDA-malleille annettiin ohjeeksi mallintaa 10 aihetta. Liian suuret aihemäärät vaikeuttaisivat tuloksena syntyvien aiheiden nimeämistä niitä kuvaaviin termein sekä aiheiden erottamista toisistaan (Fu et al. 2018). Jotta lopputuloksena syntyvä aihemalli onnistuisi mahdollisimman hyvin, aihemallin muodostamista kokeiltiin useaan otteeseen eri parametreilla, jolloin aineistolle sopivimmat parametrit löytyisivät (taulukko 9). Näitä parametrejä ovat tässä työssä iteraatiot (engl. *iterations*) ja läpikäynnit (engl. *passes*), joiden optimaaliseksi määräksi koko Helsingin kattavassa aineistossa 10 aiheelle osoittautui 30 läpikäyntiä ja 1400 iteraatiota, jonka jälkeen niiden määrien lisääminen ei enää parantunut koherenssipisteitä. Myös suuri aihemäärä nostaa koherenssipisteitä, mutta samalla se hankaloittaa aiheiden tulkintaa, sillä toisiaan lähekkäin olevien aiheiden semanttinen erottelu on haastavaa. Erottelun vaikeus korostuu tässä työssä Instagram-aineiston vuoksi, joka kuvien 33 ja 34 sanapilvien mukaisesti on varsin positiivissävytteistä. Koherenssipisteytys arvioi mallin tuloksien koherenttiutta ihmisen näkökulmasta ja on hyvä keino arvioida aihemallinnuksen onnistumista (Deshpande

2018), aihemallin onnistumista voi myös arvioida perpleksisyysarvolla, mutta se ei ole välttämättä kovinkaan luotettava (Fu et al. 2018). Valituilla parametreilla luotujen lopullisten aihemallien muodostaminen vei keskimäärin 40 minuuttia koko Helsingin kattavia aineistoja kohden ja noin 30 minuuttia yhteensä kohdealueittain.

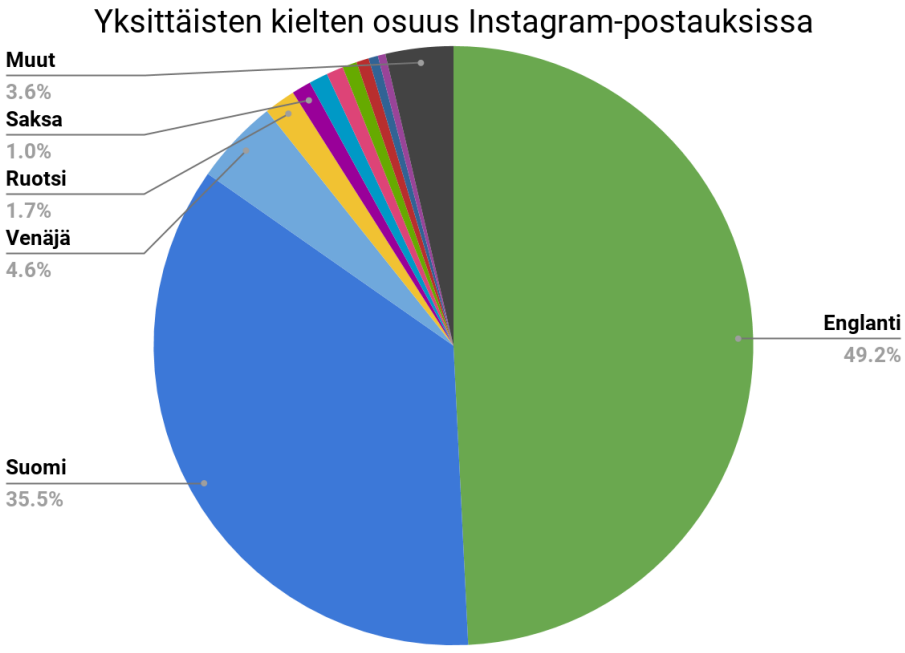
Kohdealuekohtaisessa mallinnuksessa käytettiin hieman muokattuja parametreja, koska alueet ovat pieniä ja sisältävät huomattavasti pienemmän määrän julkaisuja kuin koko Helsingin sisältävä aineisto. Aluekohtaisissa malleissa läpikäyntien määrä kaksinkertaistettiin, aiheiden määrä tiputettiin viiteen ja mallinnus tehtiin myös pelkästään adjektiiveille sekä pelkästään substantiiveihin ja verbeihin tiivistetyistä kuvateksteistä. Pelkästään substantiiveihin ja verbeihin keskittyvässä mallinnuksessa pyrittiin samanaikaisesti louhimaan aineistosta julkaisuihin liittyviä aktiviteetteja ja minimoimaan lukuisten positiivisten adjektiivien vaikutus aiheiden luomisessa. Esimerkiksi aihe, joka koostuisi pelkästään sanoista “ihana”, “hyvä” ja “lounas” voisi olla vaikea erottaa aiheesta, joka koostuu pelkästään sanoista “upea”, “mahtava” ja “ruoka”. Mallien muodostamisen kestoon vaikutti erityisen paljon läpikäyntien määrä, kun taas iteraatioiden määrällä ei ollut huomattavaa vaikutusta mallien muodostamisen kestossa. Taulukossa 9 esitelty 90 läpikäynnin mallin muodostaminen kesti lähes kaksi tuntia. Mallien muodostamisessa voi käyttää myös muita parametreja, kuten niin sanottuja esiparametreja, alphaa ja betaa. Niiden käytöllä voi parantaa mallia, mutta vaatii melko hyvää ennakkokäsitystä aineistossa esiintyvistä aiheista ja niiden jakautumisesta dokumenttien välillä ja sisällä. Ilman tätä tietoa poikkeaminen alphan ja betan oletusasetuksista ei ole suotavaa (Axelbrooke 2018). Tässä työssä käytettiin alpha- ja beta-oletusasetuksia.

3.0 Analyysi

3.1 Kielentunnistuksen tulokset

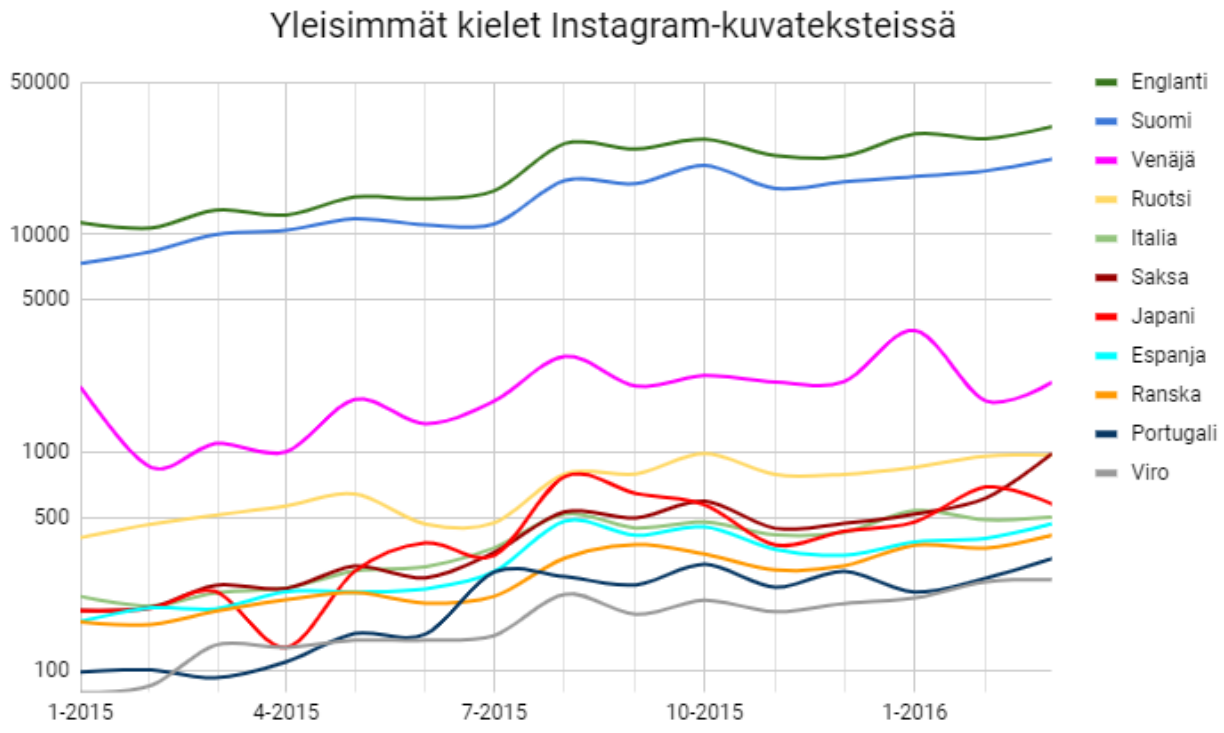
Kielentunnistuksen tuloksia tarkastellessa pelkästään aineistosta paljastuu mielenkiintoisia piirteitä kuten käytettyjen kielten hierarkia ja muutokset siinä ajassa, sekä esimerkiksi suomen ja englannin alueellinen dominanssi. Näitä tuloksia on eritelty tässä kappaleessa siten, että yleispiirteisistä tuloksista siirrytään kohti yksityiskohtaisempia englannin ja suomen kielillä tehtyjä julkaisuja.

Pelkästään yhdellä kielellä tehtyjä julkaisuja tarkastellessa kymmenen yleisimmän osuudet ovat nähtävissä kuvassa 17. Englanti ja suomi ovat ylivoimaisesti suurimmat kieliryhmät, jonka jälkeen venäjä on kolmanneksi yleisin, ruotsin ollessa vasta neljäs. Pelkästään suomeksi tai englanniksi tehtyjä julkaisuja oli yhteensä 524 487, joista 304 690 oli englanninkielisiä (49,2 %) ja 219 797 suomenkielisiä (35,5 %). Suomen jälkeen seuraavaksi suurimmat yksittäiset kielet olivat venäjä (28 337), ruotsi (10 482), saksa (6445) ja japani (6295). Venäjän lähes kolminkertainen yleisyys liittynee suomenvenäläisiin ja Venäjältä tulleisiin matkailijoihin, eikä suomenruotsissa ole tavatonta sekoittaa suomea ja ruotsia keskenään, jonka myötä on mahdollista, että osa näistä julkaisuista on täten rajautunut yhdellä kielellä tehtyjen julkaisujen aineiston ulkopuolelle.



Kuva 17. Kuvaaja yksittäisten kielten osuuksista. Japanin jälkeen tulevat nimeämättömät kielet ovat suurusjärjestyksessä: italia, espanja, ranska, portugali ja viro.

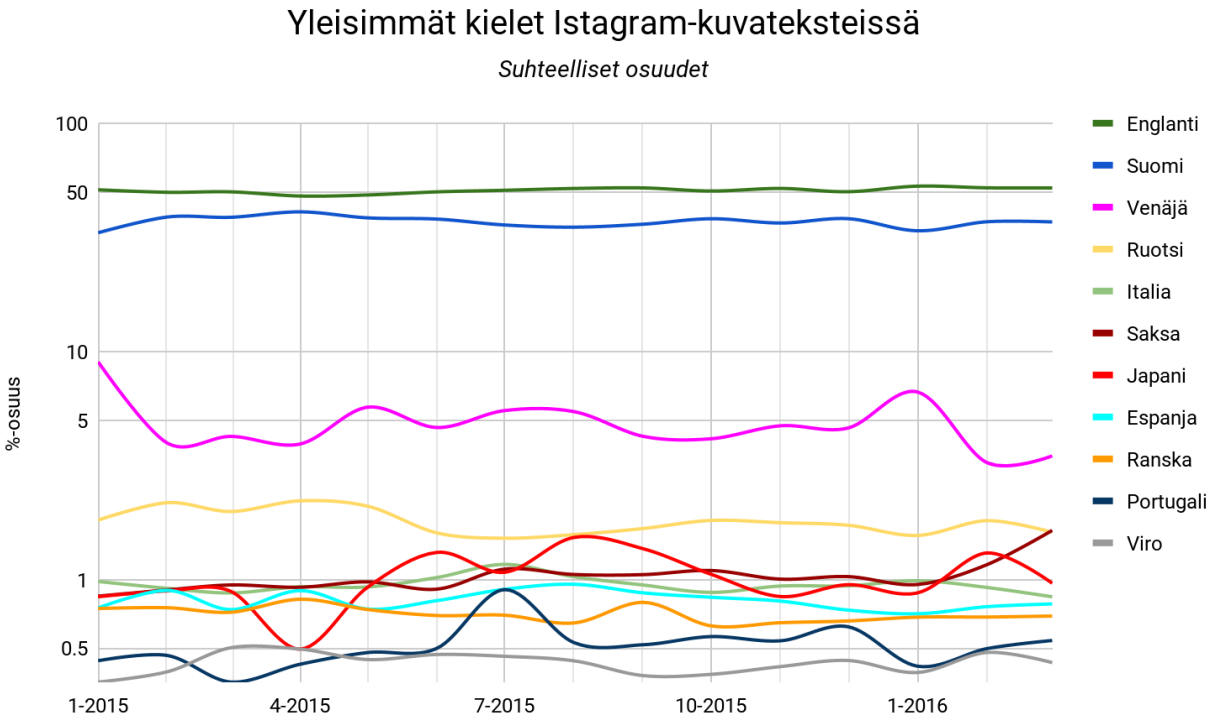
Suurin ryhmä, jossa käytettiin kahta tai useampaa kieltä, oli suomea ja englantia sekoittava ryhmä (13 471), jonka jälkeen tapahtui jyrkkä julkaisumäärien putoaminen seuraavaan kieliä sekoittavaan ryhmään, joka oli yllättävä: suomi ja italia (1802). Suomea ja ruotsia sekoittava ryhmä oli yllättävän pieni, vain 596 julkaisua, eli neljänneksitoista yleisin useampaa kuin yhtä kieltä käyttänyt ryhmä. Näin ollen on mahdollista, että suomenruotsalaiset Instagram-käyttäjät tekevät julkaisuja käyttäen vain yhtä kieltä julkaisua kohden tai koodin vaihto on jäänyt näkymättömiin, mikäli virkkeen sisällä tapahtuu kielen vaihtelua. Tällaisissa tilanteissa automaattinen kielentunnistus tunnistaa vain yhden kielen ja alhaisella varmuudella. Tätä piirrettä voisi käyttää menetelmänä kahta- tai useampaa kieltä käyttävien julkaisujen löytämiseen.



Kuva 18. Kielitunnistetun aineiston 11 suosituimman kielen julkaisut koko aineiston kattavassa aikajärjestyksessä. Kaikkien kielten osalta näyttää tapahtuneen määrällistä nousua, joka kielii Instagramin suosion kasvusta aineiston kattamalla aikavälillä. Neljä suosituinta kieltä ovat säilyttäneet asemansa koko tarkasteluvälillä, tosin japani nousi ruotsin kanssa lähes tasoihin elokuussa 2015.

Kielien hierarkkista asemaa aikasarjana tarkastellessa (kuva 18), voi huomata, että kaikkien kielien osalta tapahtuu määrällistä kasvua ja muutamia muutoksia hierarkkisessa asemassa. Kuvan asteikko on logaritminen, koska muuten muutokset suomea ja englantia pienemmissä kielissä jäävät näkymättömiin. Kaikilla kuvassa näkyvillä kielillä tehtyjen julkaisujen määrällinen kasvu kertoo Instagram-alustan suosion kasvusta. Määrällinen kasvu on suurinta julkaisuissa, jotka ovat tehty englanniksi ja suomeksi.

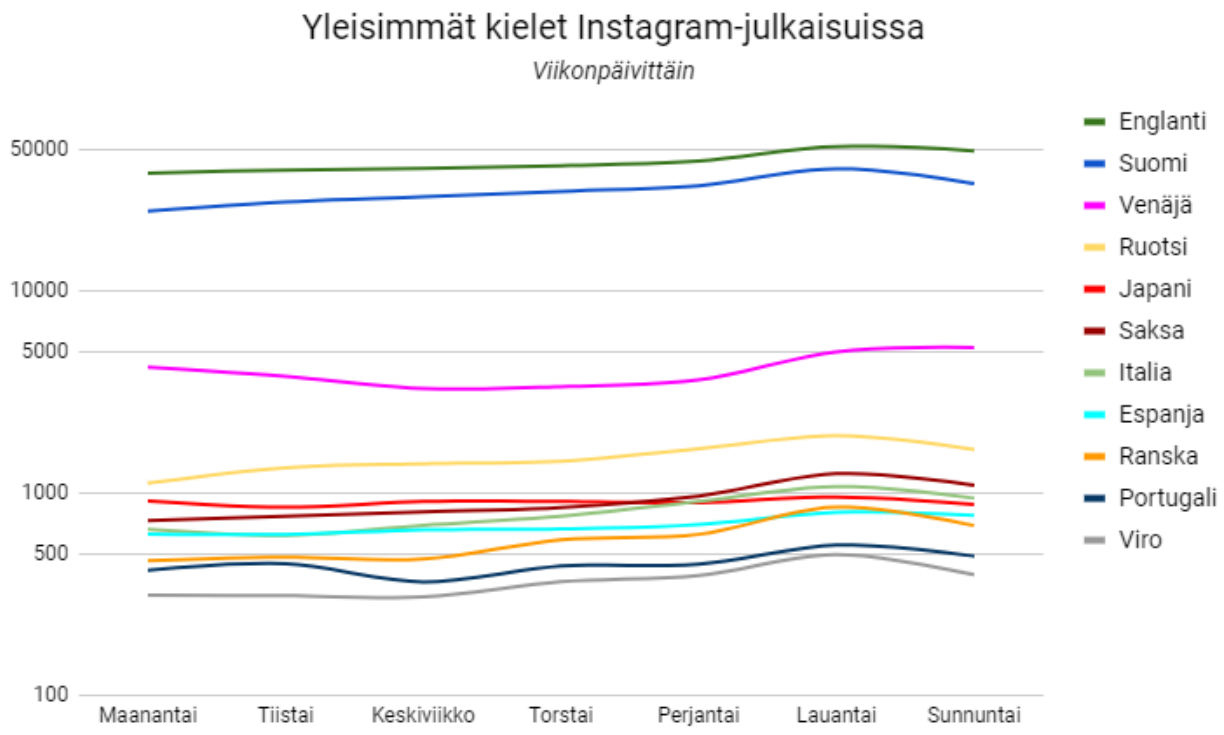
Mielenkiintoista kielien hierarkian kannalta on, että aikasarjassa kolme yleisintä kieltä säilyttävät asemansa koko aikasarjan lävitse. Ruotsi pysyy pitkään neljänneksi yleisimpänä kielenä, tosin japani pääsee lähelle ruotsia elokuussa 2015, mutta lopulta saksa ohittaa ruotsin maaliskuussa 2016 muutamalla kymmenellä Instagram-julkaisulla. Italia on aikasarjan alussa viidenneksi yleisin, mutta putoaa lopussa seitsemänneksi. Venäjä ja japani näyttävät ailahtelevan aikasarjan aikana eniten, joka on huomattu myös muissa Helsinkiin keskittyvissä sosiaalista mediaa hyödyntävissä tutkimuksissa (Hiippala et al. 2018). Ailahtelun syyksi on arveltu olevan venäjän kielen osalta Ukrainan konfliktista kumpuavat Venäjän valtion vastaiset talouspakotteet ja japanin osalta mobiililaitteiden erittäin suuri yleisyys (Hiippala et al. 2018).



Kuva 19. Yleisimpien Instagramissa Helsingin alueella käytettyjen kielien suhteellinen osuus aineistossa vuoden 2015 tammikuusta vuoden 2016 maaliskuun loppuun.

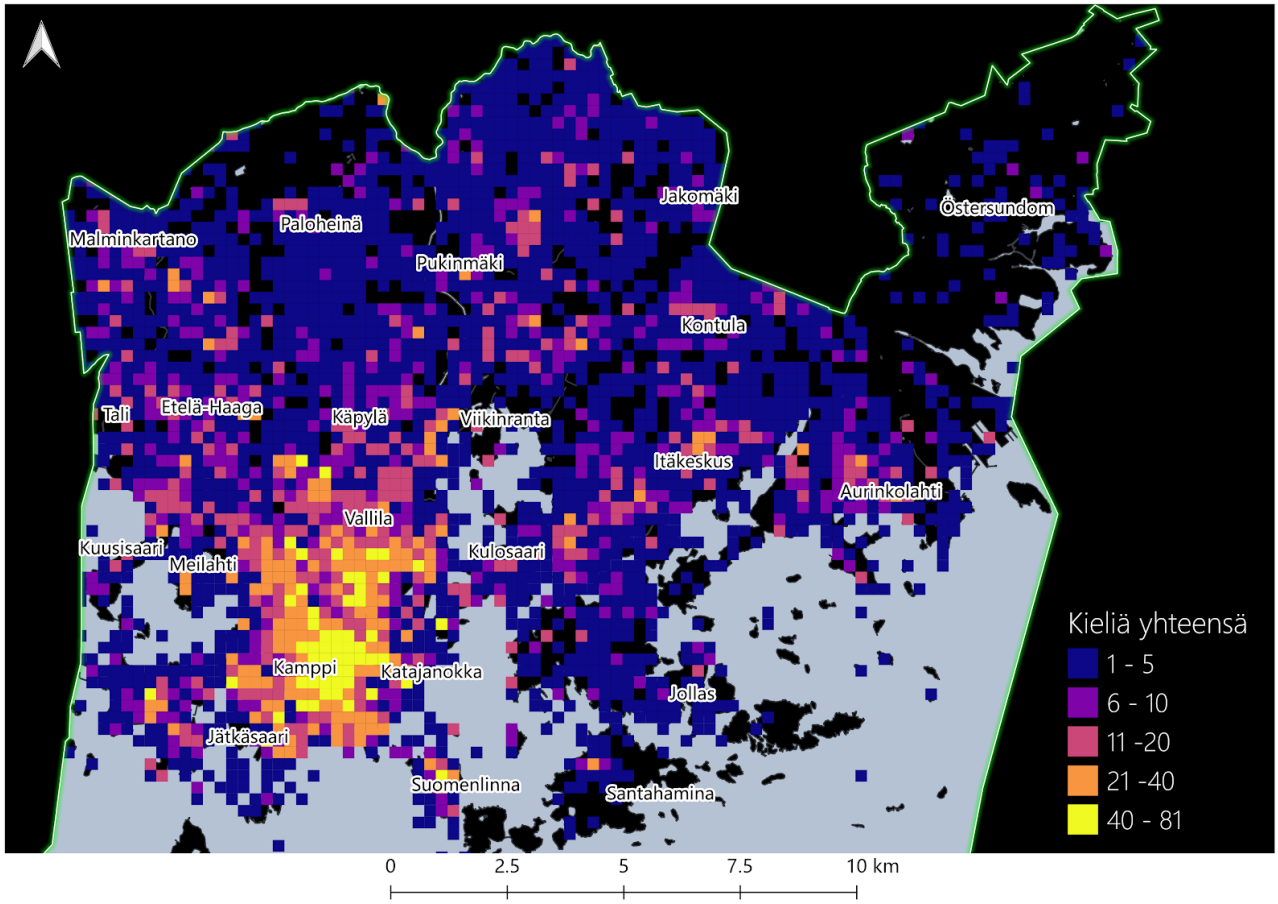
Tarkasteltaessa eri kielillä tehtyjen julkaisujen suhteellisia osuuksia Helsingin alueen Instagram-aineistosta paljastuu uusia piirteitä (kuva 19). Vaikka julkaisujen määrä on ollut jatkuvasti kasvussa, englannin ja suomen suhteellisissa osuuksissa ei ole juurikaan tapahtunut muutoksia: englanti on pysytellyt noin 50 % osuudessa kaikista julkaisuista koko aineiston kattavan aikavälin ajan, suomen osuus kasvoi aivan alussa, mutta on sen jälkeen pysytellyt 38 % osuudessa läpi aineiston. Pienimmissä Instagram-kielissä taas on tapahtunut suhteellisia muutoksia, esimerkiksi venäjän osuus on pienentynyt vuoden 2015 alusta

joulukuuta lukuun ottamatta jatkuvasti. Syynä saattaa olla Venäjän valtion vastaiset talouspakotteet, joihin myös Suomi on osallistunut. Japanin suhteellinen osuus on sahannut edestakaisin, saavuttaen korkeimman osuutensa tarkastelujakson keskivaiheilla kesä-syyskuussa 2015. Saksan osuus on pysytellyt samana, 1 % lähettyvillä, mutta nousi maaliskuussa 2016 ylöspäin lukemiin 1,6 %.



Kuva 20. Yleisimmät Instagramissa käytetyt kielet Helsingin alueella viikonpäivittäin.

Kun yleisimpiä Instagram-kieliä visualisoidaan viikonpäivittäin aineistossa ei näy suuria muutoksia. Lähes kaikki kielet pitävät hierarkkisen asemansa lukuun ottamatta saksaa ja ranskaa. Saksa nousee lauantaisin ruotsin jälkeen viidenneksi yleisimmäksi kieleksi japanin ohi. Japanin ohittaa viikonloppuisin myös italia. Ranska ohittaa lauantaisin espanjan. Tämä piirre liittynee pitkälti viikonloppumatkoihin Helsinkiin näistä Euroopan maista, mutta myös osittain siihen, että viikonloppu on ylipäätään aktiivisinta Instagram-julkaisemisen aikaa (kuvat 8 ja 9). Japani on ainoa selkeästi ei-eurooppalainen kieli, joka kielinee matkailun vaikutuksesta. Ei ole yllättävää, että japaniksi tehtyjen julkaisujen määrä pysyy samana läpi viikon, sillä esimerkiksi Japanista ei todennäköisesti lähdetä Helsinkiin lyhyelle viikonloppumatkalle vaan pidemmäksi ajaksi, jolloin viikonlopun merkitys ei korostu.

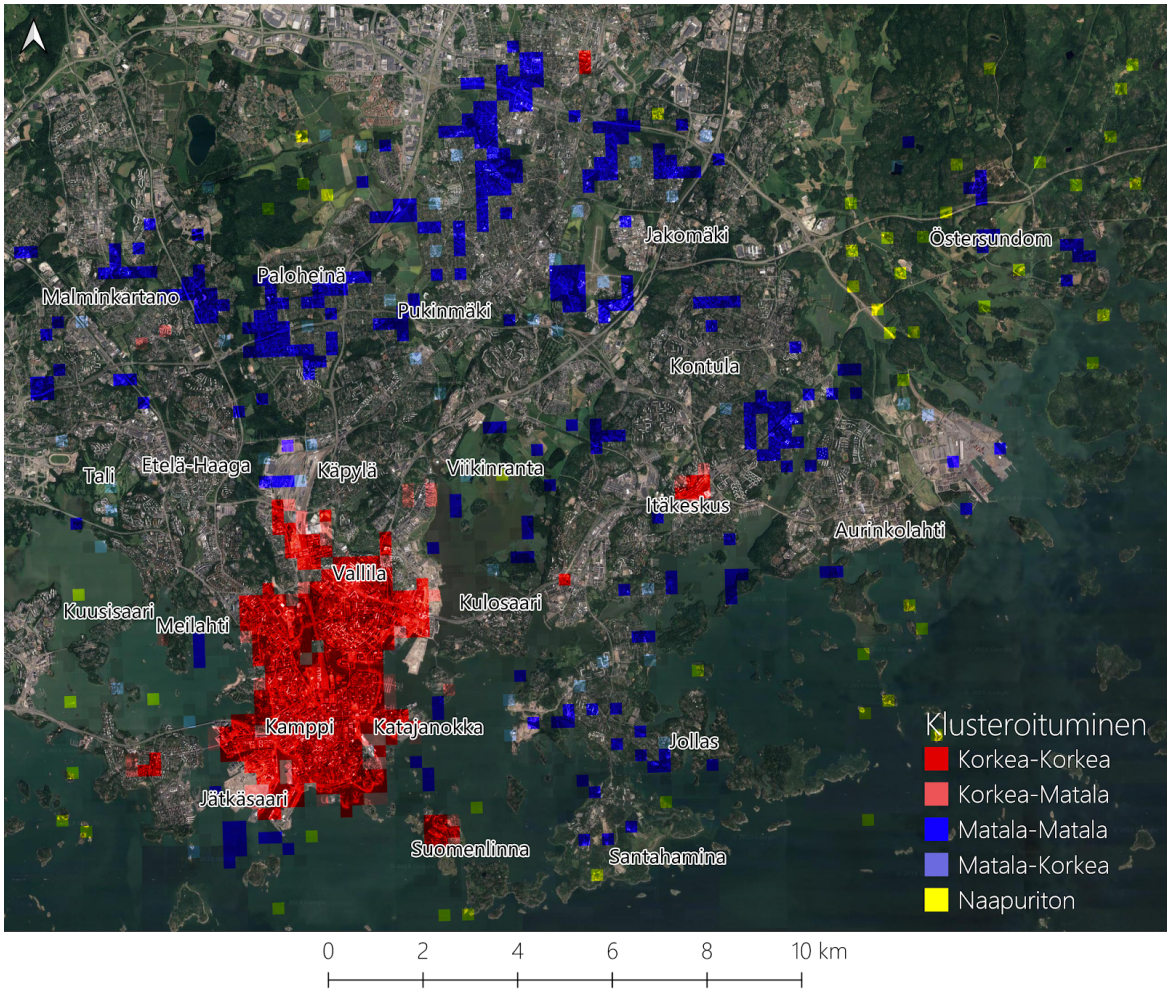


Kuva 21. Tunnistettujen kielten määrät eli kielellinen rikkaus visualisoituna YKR-ruudukkoon. Kielellinen rikkaus näyttää seuraavan Helsingin yhteiskuntarakennetta.

Kun aineistosta tunnistetut käytetyt kielet koostetaan yleisesti käytössä olevaan 250 x 250 metrin YKR-ruudukkoon kartalle (kuva 21) on nähtävissä, että kielellisesti rikkaimmat alueet keskittyvät ydinkeskustaan, Kallioon, Suomenlinnaan, Hartwall Arenaan ja Lauttasaareen. Matalampia kielellisen rikkauden keskittymiä on nähtävissä itäisen metroradan varrella, Arabianrannassa, Pukinmäen lähettyvillä ja Munkkiniemessä. Kielimäärien alueellista rakennetta voi selittää sen linkittyminen alla olevaan yhteiskuntarakenteeseen. Korkeat käytettyjen kielten määrät voivat myös viestiä siitä, mitkä alueet ovat kansainvälisesti näkyviä alueita Helsingissä ja mitkä eivät. Hieman yllätyksettömästi, keskustasta etäämmällä olevat alueet eivät juurikaan näyttäydy kansainvälisinä alueina verrattuna hyvin saavutettavissa oleviin alueisiin, jotka sijaitsevat hyvien kulkuyhteyksien varrella.

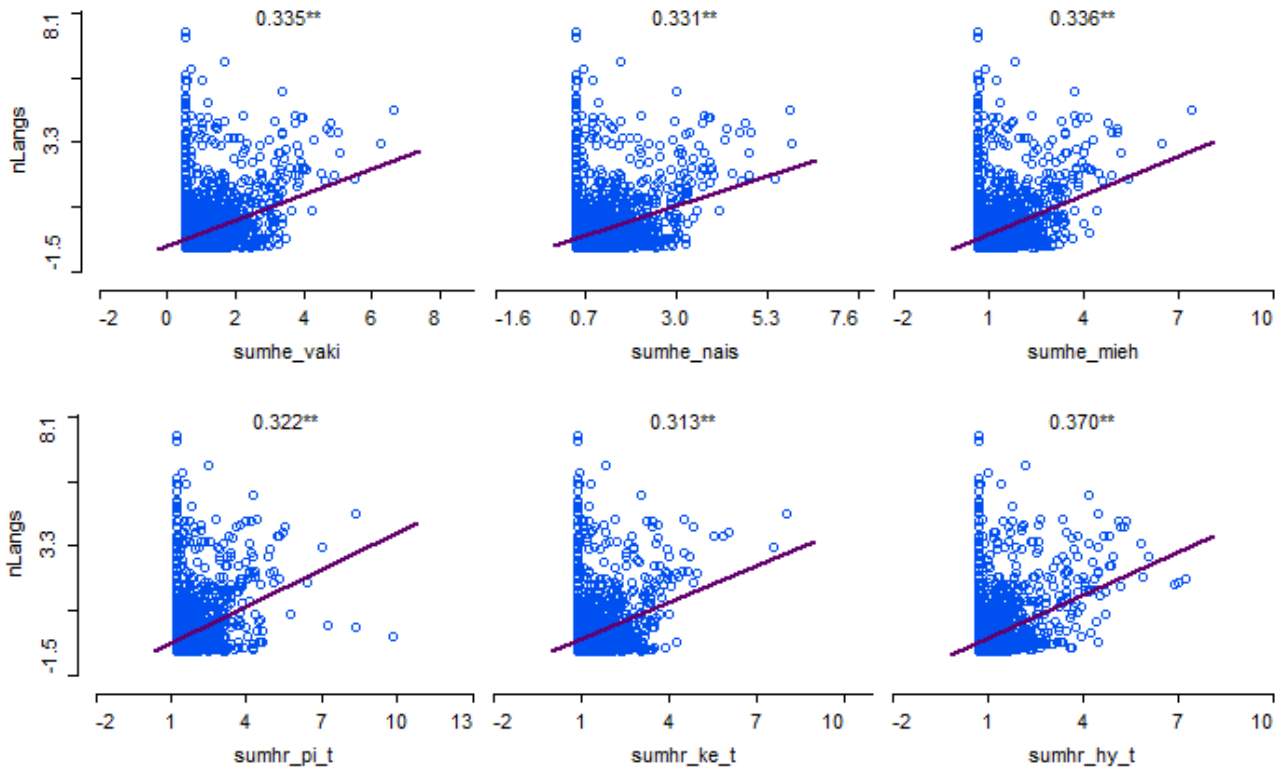
Kielellisen rikkauden spatiaalista autokorrelaatiota tarkastellessa kuvasta 22 Helsingin kantakaupunki erottuu selkeänä korkean kielellisen rikkauden klusterina. Klusterointi on laskettu paikallisella Moranin indeksillä, käyttäen Queen-naapuruutta ja 0.05:n merkitsevyystasoa. Klusterointi kertoo

Instagram-julkaisuissa näkyvästä digitaalisen kaupunkitilan kielimaiseman spatiaalisesta rakenteesta. Kantakaupungin muodostaman suuren klusterin ulkopuolella sijaitsevia korkeiden kielimäärien klustereita ovat Itäkeskus, Lauttasaari, Suomenlinna ja Hartwall Arenan lähialueet. Isoja pienien kielimäärien klustereita on lähinnä Pohjois-Helsingissä, joissa julkaisujen määrät ovat muutenkin pieniä, mutta myös Mellunkylän alueella Itä-Helsingissä, sekä Jätkäsaaren eteläkärjessä. Muuten pienet kielimäärät vaikuttavat muodostavan yksittäisiä klustereita, jotka ovat toisistaan erillään. Ympäröivistä soluista selkeästi poikkeavia arvoja (Korkea-Matala tai Matala-Korkea) on melko vähän ja ne seurailevat pitkälti isompia klustereita. Tosin Kumpulan ja Arabianrannan alueella on poikkeuksellinen solujoukko, jossa kyseiset solut pitävät sisällään korkean määrän kieliä, mutta ympäröivät sisältävät pienen määrän kieliä. Tämän taustalla saattaa olla alueella sijaitsevat korkeakoulukampukset, joissa opiskelee ja työskentelee melko suuri määrä ulkomaalaistaustaisia henkilöitä sekä matkailijoita houkutteleva Arabian outlet-myymälä.



Kuva 22. Käytettyjen kielien määrän klusteroituminen paikallisella Moranin I:llä. Klustereiden muodostamisessa käytettiin Queen-naapuruutta ja näkyvien klustereiden tilastollinen merkittävyys on 0.05.

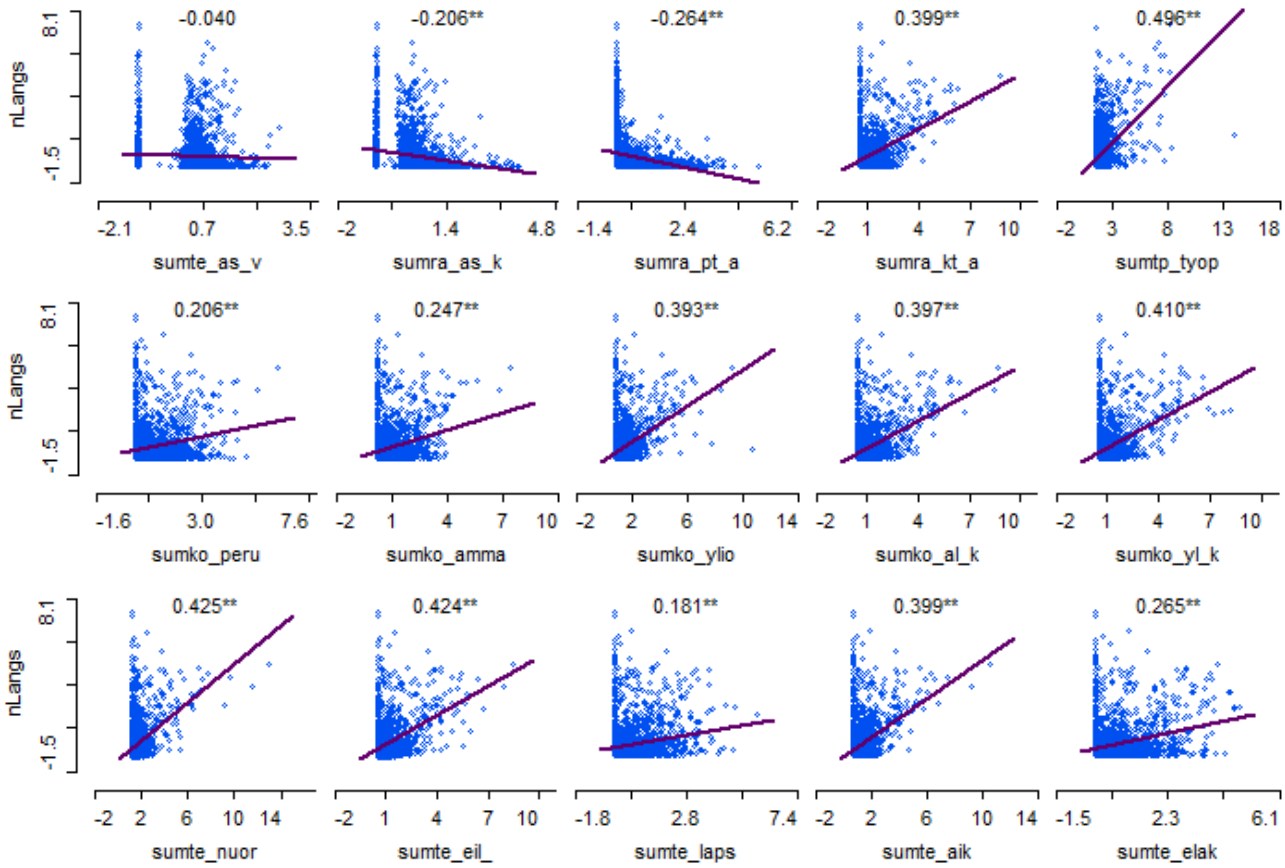
Käytettyjen kielien määrien klusterointi ei varsinaisesti yllätä, mutta Itäkeskus ja Lauttasaari ovat mielenkiintoisia pienempiä klustereita. Itäkeskuksen korkea kielimäärää selittänee suurien ostoskeskuksien sijaitseminen alueella ja se, että itäinen metrorata haarautuu Itäkeskuksesta itään päin, muodostaen siitä liikenteen solmukohdan. Suuret ostoskeskukset ja liikenteen solmukohdat ovat usein alueita, joilla liikkuu suuri määrä ihmisiä. Lisäksi itäinen Helsinki on hieman vahvemmin ulkomaalaistaustaisten asuttama kuin muut osat Helsinkiä. Lauttasaaren klusteri on hieman vaikeampi selittää, mutta sekin on eräänlainen liikenteen solmukohta Helsingin ja Espoon välillä. Pienen kielellisen rikkauden klustereita vaikuttaa olevan pitkälti pientalovaltaisilla alueilla kuten Vartioharjussa, Paloheinässä ja Siltamäessä. Lisäksi vähäisen käytettyjen kielten lukumäärien alueita on myös luonnonalueilla kuten metsissä ja merellä, joka ei ole yllättävää.



Kuva 23. Kaksi standardoitua scatterplot-matriisia kielten lukumäärän ja usean YKR-ruudukon (SYKE 2016) muuttujan kanssa. Matriisista on nähtävissä, että Pearsonin korrelaatiot ovat heikkoja, mutta tilastollisesti merkittäviä ($p < 0.01$). Ylemmällä rivillä ruudussa asuvien miesten lukumäärä korreloi vahvimmin käytettyjen kielten määrän kanssa. Alemmalla rivillä hyvätuloisiksi luokitellut korreloi vahvimmin käytettyjen kielten määrän kanssa.

Tarkastellessa kuvaa 23, näkee kielten lukumäärän korreloivan heikosti asukasmäärän ja tuloluokkien kanssa. Nämä tulokset eivät yllätä, sillä julkaisujen lukumäärät eivät korreloineet juuri lainkaan väestöllisten

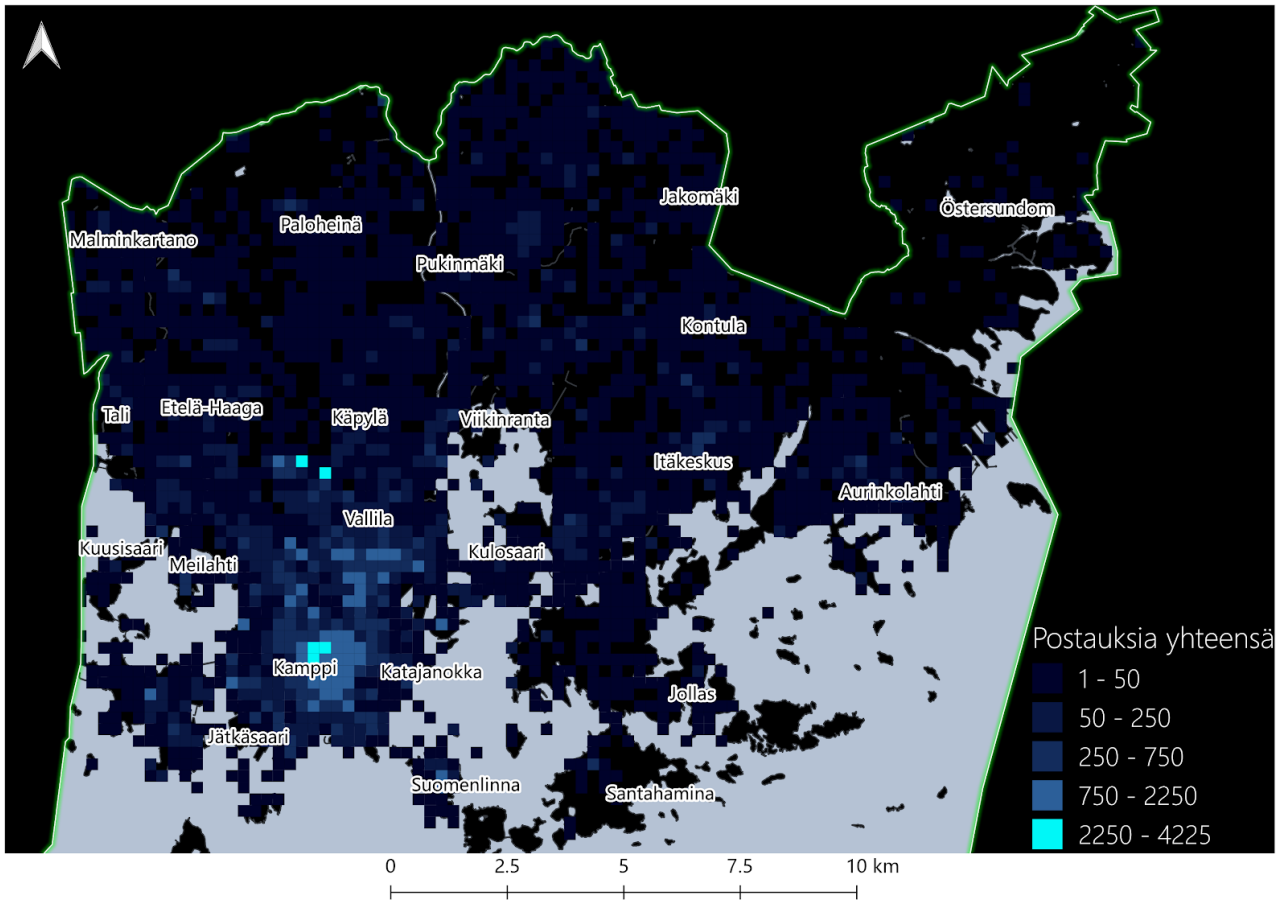
muuttujien kanssa kuvassa 15. Hieman mielenkiintoisempia havaintoja saadaan, kun julkaisuissa käytettyjen kielten lukumäärää verrataan muihin YKR-ruudukon muuttujiin, kuten on tehty seuraavaksi kuvassa 24.



Kuva 24. Kolme standardoitua scatterplot-matriisia käytettyjen kielten ja usean YKR-ruudukon muuttujan kanssa. Ylimmällä rivillä on asuntoihin liittyviä muuttuja, keskimmäisellä rivillä koulutustasoon liittyviä muuttujia ja alimmalla tasolla talouksien tyyppien lukumääriä. Kaikki Pearsonin korrelaatiot, asumisväljyyttä lukuun ottamatta, ovat tilastollisesti merkittäviä ($p < 0.01$).

Kuvasta 24 on nähtävissä hieman mielenkiintoisempia korrelaatioita, kuin kuvasta 23. Esimerkiksi asumisväljyys (sumte_as_v) ja asuntojen keskipinta-ala (sumra_as_k) eivät korreloi käytettyjen kielten määrän kanssa juuri lainkaan. Kuvan 22 kartan antaman kuvan mukaisesti ei ole yllättävää, että pientaloasuntojen lukumäärällä (sumra_pt_a) on heikkoa negatiivista korrelointia käytettyjen kielten lukumäärän kanssa. Kerrostaloasunnoilla (sumra_kt_a) on keskinkertaista positiivista korrelointia ja työpaikkojen (sumtp_tyop) määrällä on vahvempaa keskinkertaista korrelointia kielten lukumäärän kanssa. Koulutustasoa katsellessa ylemmän korkeakoulututkinnon suorittaneiden määrä (sumko_yl_k) korreloi käytettyjen kielten lukumäärän kanssa vahvimmin, mutta sekin keskinkertaista heikompana. Taloustyyppien lukumääriä ja käytettyjen kielten määriä vertaillessa vahvimmin korreloi nuorten yksinasuvien talouden

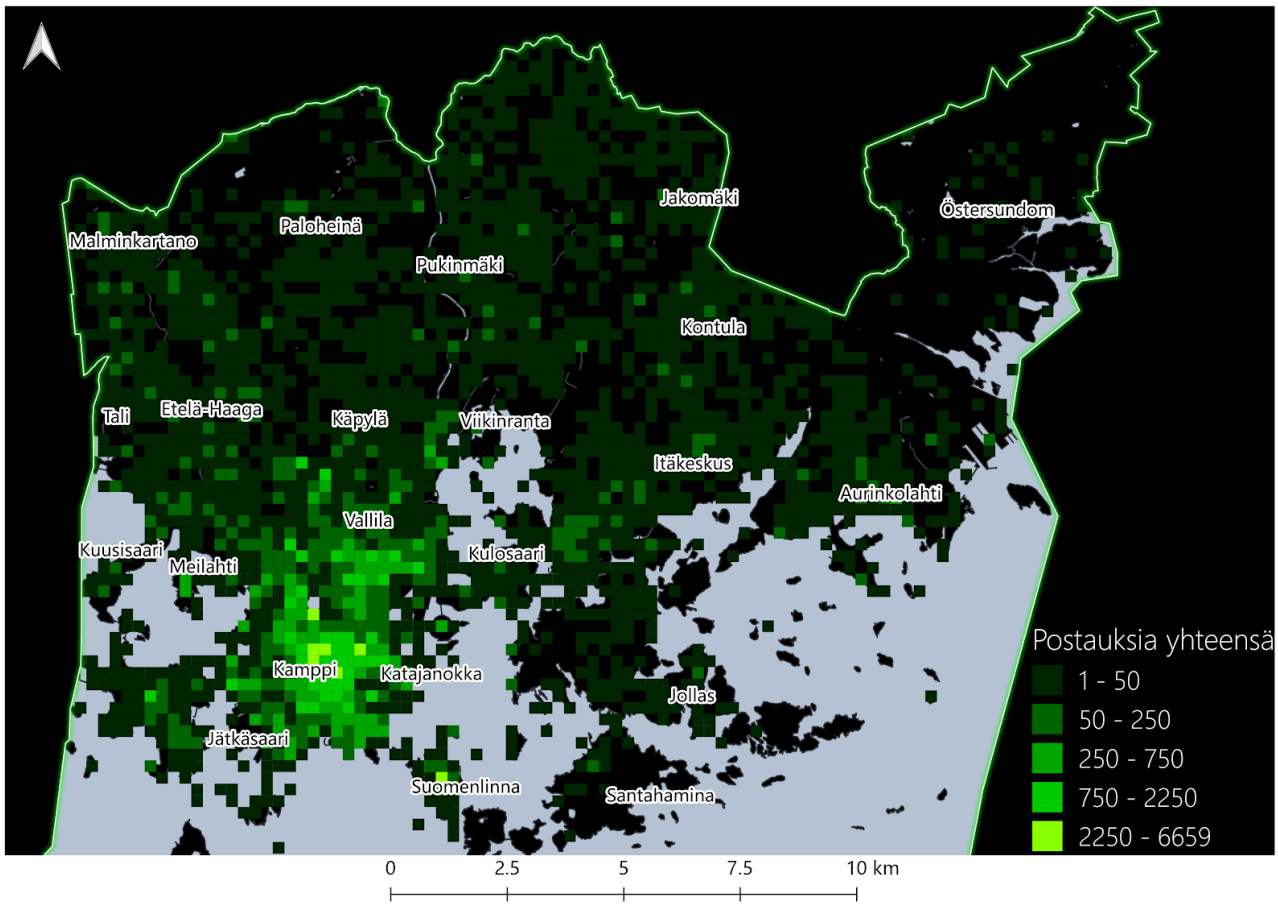
(sumte_nuor) ja nuorten lapsettomien perheiden taloudet (sumte_eil). Lapsiperheet (sumte_laps) eivät juurikaan korreloi käytettyjen kielten määrän kanssa, myös eläkeläisten taloudet (sumte_elak) korreloivat heikosti käytettyjen kielten lukumäärän perusteella. Näiden tulosten valossa käytettyjen kielten lukumäärää ei vaikuta olevan kytköksissä kovin voimakkaasti julkaisualueen sosioekonomisiin muuttujiin. Käytettyjen kielien lukumäärää ei siis voi kovin vakuuttavasti selittää näillä muuttujilla, mutta niillä voi olla jonkinlainen keskivahva kytkös joka tapauksessa. Korkea koulutusaste ja nuorten ihmisten taloudet vaikuttavat olevat parhaiten korkeiden kielimäärien kanssa korreloivia alueellisia muuttujia, joka ei sinänsä ole yllättävää sillä Instagram on erityisesti nuorten aikuisten suosima sosiaalisen median palvelu (Greenwood et al. 2016).



Kuva 25. Suomenkieliset Instagram-julkaisut Helsingin alueella. Suurin keskittymä on Helsingin ydinkeskusta, yksittäisiä keskittymiä ovat myös Hartwall Arenan lähettyvillä Vallilan ja Etelä-Haagan välissä.

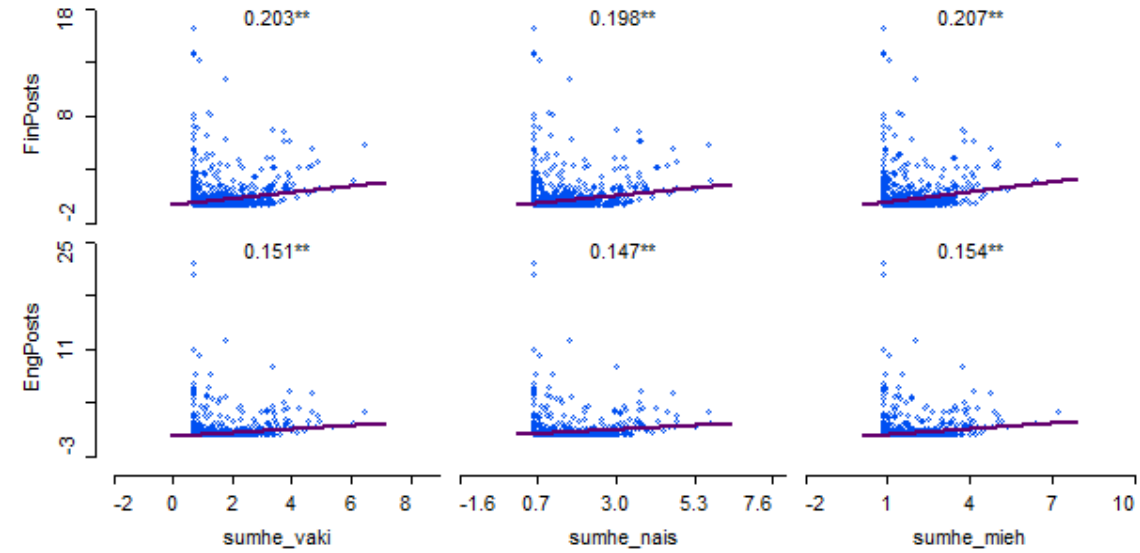
Pelkästään suomenkielisiä Instagram-julkaisuja tarkastellessa kuvasta 25 on nähtävissä selkeä keskittyminen Helsingin ydinkeskustan ja Kallion alueille, vaikkakin yksittäisiä piikkejä on nähtävissä pohjoisempana Hartwall Arenan ja Pasilan Messukeskuksen lähettyvillä. Hartwall Arenan näkyminen ei ole yllättävää, sillä se oli neljänneksi mainituin sijainti koko aineistossa (kuva 8). Muuten näyttäisi siltä, että julkaisujen määrä

seurailee asukasmääriä melko selkeästi, tosin tilastollista näyttöä tälle “seurailulle” on vain heikosti (kuva 27).

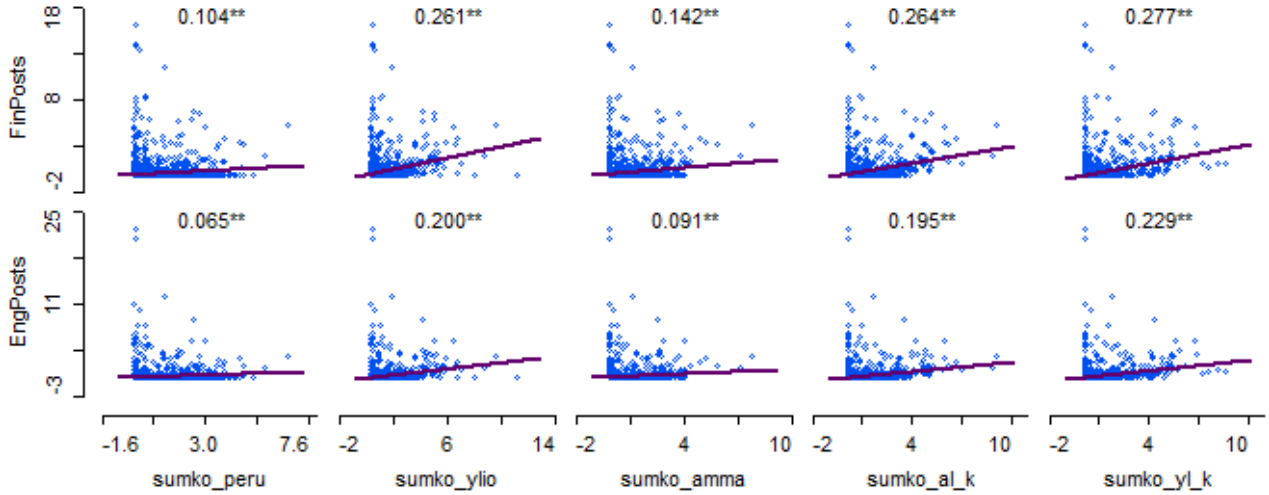


Kuva 26. Englanninkielisten Instagram-julkaisujen alueellinen levinneisyys Helsingissä.

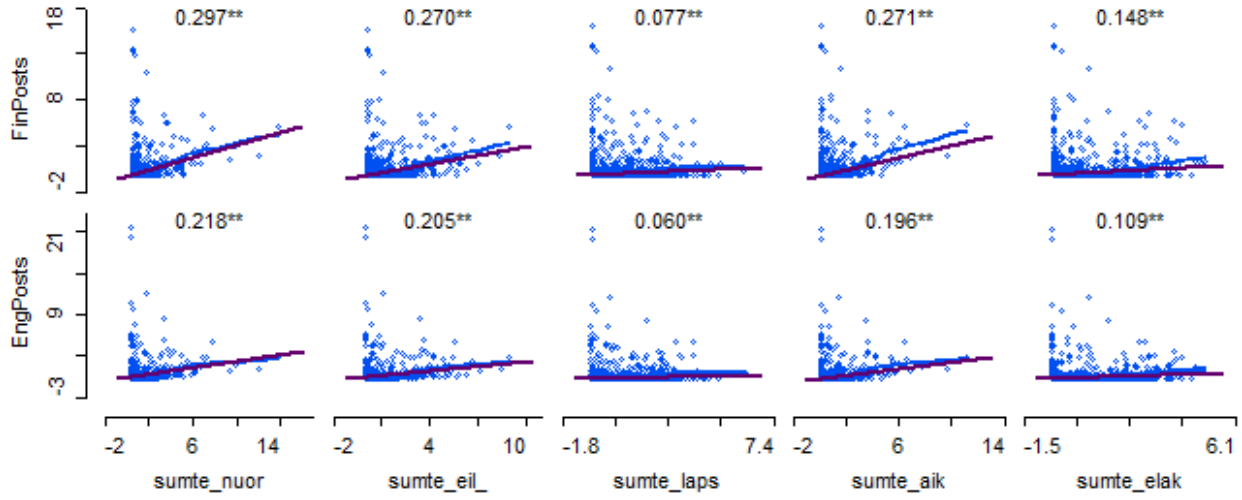
Englanninkielisten Instagram-julkaisujen osalta kuvassa 26 on nähtävissä samankaltaista alueellista rakennetta kuin suomenkielisen aineiston kanssa (kuva 25), tosin Hartwall Arena ei ole yhtä selkeä keskittymä englanninkielisissä julkaisuissa. Selkeänä erona ovat myös kantakaupungin ulkopuoliset alueet, joilla ei ole yhtä paljon julkaisuja englanniksi kuin suomeksi. Suomenlinna on myös englanninkielisessä aineistossa selkeämpi, ydinkeskustan kaltainen keskittymä kuin suomenkielisessä aineistossa. Englanninkielisissä julkaisuissa on mahdollisesti enemmän matkailuun liittyvää aihesisältöä, joka selittäisi melko hyvin Suomenlinnan ja eteläisen ydinkeskustan selkeämpää näkymistä tässä aineistossa verrattuna suomenkielisiin aineistoon.



Kuva 27. Standardoitu scatterplot-matriisi suomeksi ja englanniksi tehdyistä julkaisuista sekä YKR-ruudukon väkimäärään liittyvistä muuttujista. Pearsonin korrelaatiot ovat tilastollisesti merkittäviä ($p < 0.01$), mutta heikkoja.

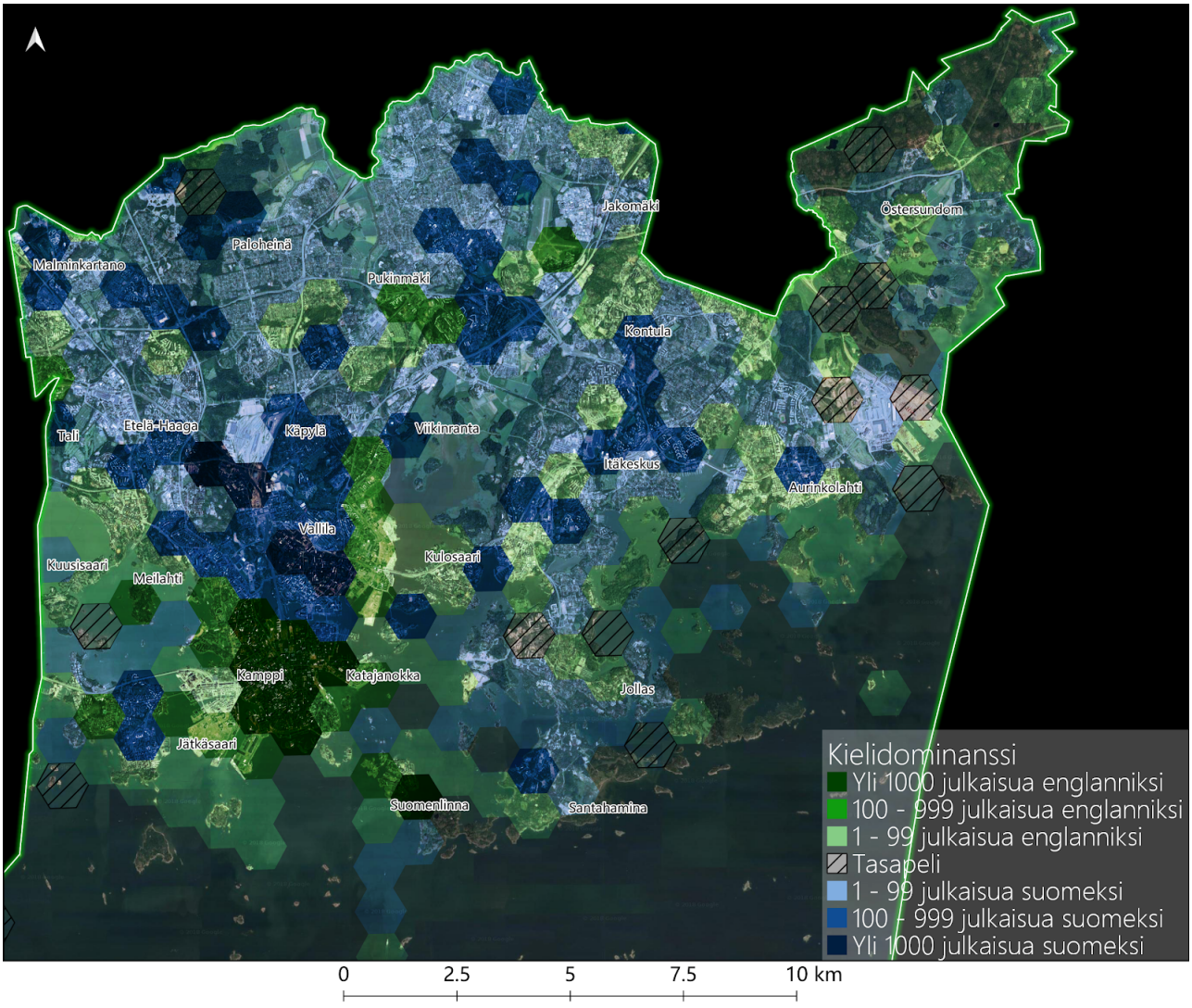


Kuva 28. Standardoitu scatterplot-matriisi suomeksi ja englanniksi tehdyistä julkaisuista sekä YKR-ruudukon koulutukseen liittyvistä muuttujista. Kummallakin kielellä julkaisujen Pearsonin korrelaatio koulutukseen liittyvien muuttujien kanssa ovat tilastollisesti merkittäviä ($p < 0.01$), mutta heikkoja.



Kuva 29. Standardoitu scatterplot-matriisi suomeksi ja englanniksi tehdyistä julkaisuista sekä YKR-ruudukon kotitalouksiin liittyvistä muuttujista. Pearsonin korrelaatiot ovat tilastollisesti merkittäviä ($p < 0.01$), mutta melko heikkoja.

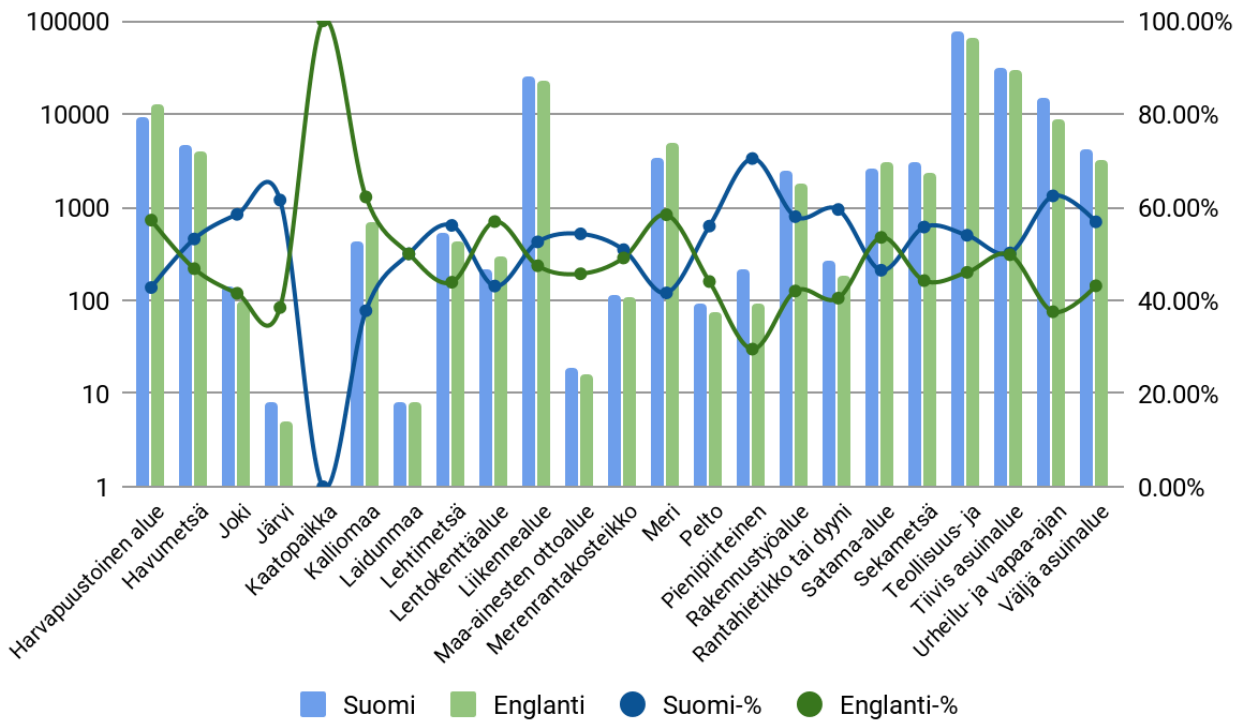
Kuvia 27, 28 ja 29 tarkastellessa huomaa, että suomenkieliset ja englanninkieliset julkaisut korreloivat heikoimmin samojen YKR-muuttujien kanssa, joiden korrelaatiota testattiin käytettyjen kielten lukumäärän kanssa (kuvat 23 ja 24) ja julkaisujen kokonaismäärän kanssa (kuva 16). Heikommista korreloinneista huolimatta ne ovat tilastollisesti merkittäviä ja pitävät sisällään samaa logiikkaa kuin käytettyjen kielten määrien korrelaatiot: korrelaatio on vahvinta korkeasti koulutettujen ja nuorten talouksien alueilla. Kuvien 22 - 28 tarkasteluista paljastuneet piirteet tukevat ennakkokäsitystä siitä, että Instagram-alustalla tapahtuvat julkaisut keskittyvät positiivisen ja henkilökohtaista menestystä viestivän sisällön lisäksi myös alueille, joilla on paljon palveluita ja korkeat neliöhinnat. Samankaltaisia havaintoja Instagram-julkaisuista ja käyttäjistä on tehty myös muualla (Sheldon & Bryant 2016; Keipi et al. 2018).



Kuva 30. Täysin suomeksi ja englanniksi tehtyjen postauksien dominanssin spatiaalinen rakenne yhden kilometrin kennostossa. Suomen kieli dominoi ydinkeskustan ulkopuolisia alueita, erityisesti Kalliota, kun taas englannin kieli on vahvempaa ydinkeskustassa.

Suomen- ja englanninkielisten julkaisujen välisiä eroja tarkastellessa alueellisesti Helsingin digitaalisen kaupunkitilan kieleen perustuva rakenne alkaa hahmottua. Kuvasta 30 on nähtävissä englannin- ja suomenkielisten julkaisujen lukumääriin perustuva alueellinen dominanssi Helsingissä, jossa ilmoitetaan, kuinka paljon enemmän kyseisellä kielellä julkaisuja on yhtä kennoa kohden. Kennot ovat yhden kilometrin kennoja eli vastakkaisten sivujen etäisyys toisistaan on kilometri. Yleispiirteisesti spatiaalisesta kuvioinnista voi todeta englanninkielisten julkaisujen keskittyvän rannikoiden läheisyyteen, kun taas suomenkieliset julkaisut vaikuttavat tapahtuvan vahvemmin sisämaassa. Kielellistä dominanssia on mitattu muualla myös useilla ekologiasta peräisin olevilla indekseillä, kun tarkastelussa on ollut useita eri kieliä (Peukert 2013), mutta kahden kielen väliseen vertailuun absoluuttisiin lukuihin perustuva vertailu lienee riittävä.

Suomen kieli dominoi eteläisen keskustan ulkopuolisia alueita ja taulukossa 3 todettu suomenkielisten julkaisujen yleisyys Kalliossa näkyy myös kartalla. Suomen kielen suosioon Kallion alueella syynä voi olla esimerkiksi tunnetuimpien turisteja puoleensa vetävien Helsingin nähtävyyksien sijaitseminen pääasiallisesti eteläisessä kantakaupungissa. Kallion keskittymän lisäksi toinen vahvin suomen kielen dominanssialue on Hartwall Arenan ja Ilmalan alueella. Hakaniemen, Käpylän ja Taka-Töölön muodostama alue sisältää nämä vahvat suomen kielen dominanssikennot ja on samalla alueellisesti suurin yhtenäinen samaan luokkaan kuuluva suomen kielen dominanssikennojen alue. Ydinkeskustan alueella englanti muodostaa suurimman dominanssikeskittymänsä eikä sillä juurikaan ole muita selkeitä isoja keskittymiä, vaan pelkästään yksittäinen Suomenlinnan kenno. Pienempiä englannin kielen keskittymiä on Pukinmäessä, Malmin lentokentällä ja Lautta- sekä Seurasaaressa. Saman verran julkaisuja molemmilla kielillä on vain hajanaisesti ja lähinnä alueilla, joilla ei ole suuria julkaisumääriä. Töölönlahden eteläpuolinen Helsinki vaikuttaa siis olevan pääsääntöisesti englanninkielisten julkaisujen dominanssialuetta, kun taas suomenkielisten julkaisujen dominoima alue alkaa Hakaniemiestä ja Töölönlahden pohjoispuolelta.

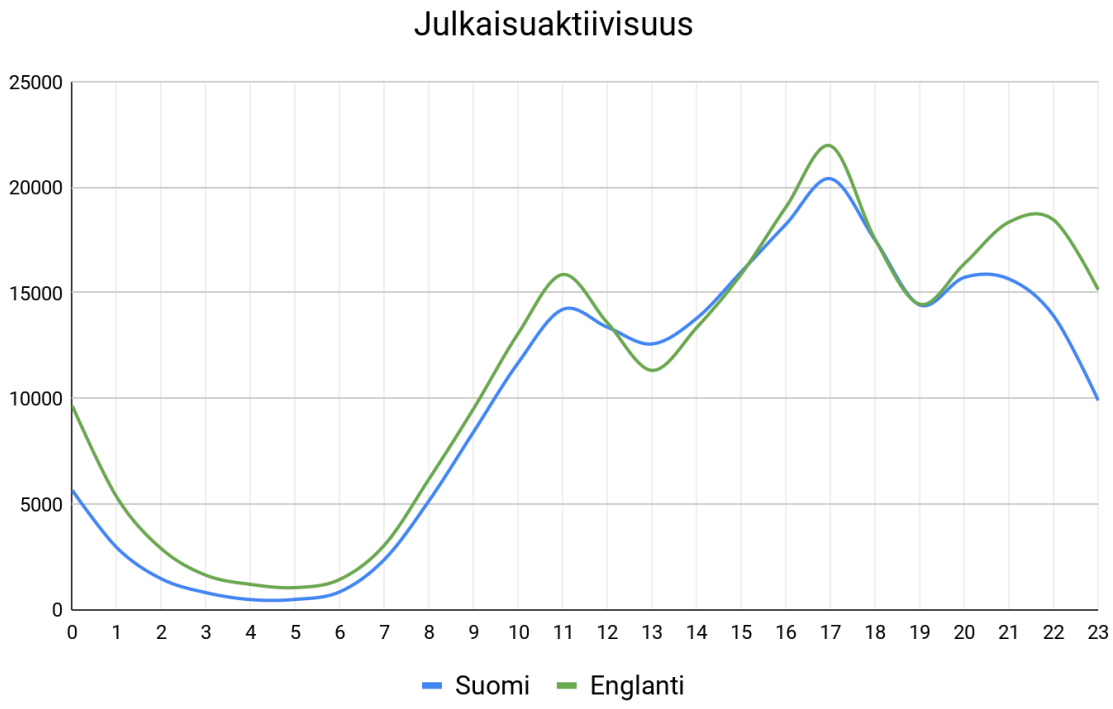


Kuva 31. Englannin- ja suomenkielisten Instagram-julkaisujen lukumäärät ja osuudet jaoteltuna vuoden 2016 CORINE-maankäyttöluokittain.

Kuvasta 31 on selkeästi nähtävissä julkaisujen keskittyminen rakennetuille alueille luonnonalueiden sijaan. Suurin maankäyttöluokka julkaisuja kohden on teollisuus- ja palvelualueiden maankäyttöluokka. Havainto on yhtäläinen kuvassa 22 todetun alueellisen korrelaation kanssa. Seuraavaksi suurimmat ovat tiivis asuinalue, liikennealue, urheilu- ja vapaa-ajan alue sekä harvapuustoinen alue. Pienimmät maankäyttöluokat ovat järvi, laidunmaa, maa-ainesten ottoalue, jotka eivät ole yllättäviä, sillä näitä on Helsingissä todella vähän. Luonnonalueista suurimmat ovat havumetsät, meri ja sekametsä.

70 % tai suuremmalla varmuudella tunnistetuista julkaisuista enemmistö on suomeksi tehtyjä, jonka vuoksi suomi on yleisempi julkaisukieli useimmissa luokissa. Englanti on suositumpi kieli vain kuudessa maankäyttöluokassa 23:sta: harvapuustoinen alue, kaatopaikka (vain 1 julkaisu), kalliomaa, lentokenttäalue, meri ja satama-alue. Lentokenttä- ja satama-alueiden vahvempi englanninkielisyys vahvistaa englanninkielisten julkaisujen kytköstä matkailuun. Kaatopaikka-maankäyttöluokan yksittäinen englanniksi tehty julkaisu ei todennäköisesti ole merkittävä, joten se sivuutetaan epäolennaisena poikkeamana. Suomi on selkeästi englantia yleisempää muutamassa maankäyttöluokassa: joki, järvi, pienipiirteinen

maatalousmosaiikki, rakennustyömaa, rantahietikko tai dyyni, sekä urheilu- ja vapaa-ajan alue. Nämä maankäyttöluokat liittyvät maatalousmosaiikkia ja rakennustyömaita lukuun ottamatta melko selkeästi vapaa-aikaan vahvistaen suomenkielisten julkaisujen vahvaa kytkeytymistä vapaa-aikaan ja heikompaa kytkeytymistä työskentelyyn. Kuvan 31 kuvaajaa tarkastellessa tulee muistaa Instagram-aineiston kohdepisteisiin perustuva sijaintirakenne, jonka vuoksi maankäyttömuotoihin luokitelluissa julkaisumäärissä voi olla pieniä virheitä.



Kuva 32. Suomen- ja englanninkielisen aineiston julkaisuaktiivisuus kellonajan mukaan.

Kuvasta 32 voi huomata Instagram-aineiston suomen- ja englanninkielisten kielien julkaisujen noudattavan keskenään samankaltaista rytmiikkaa, tosin englanninkielinen julkaisuaktiivisuus on suurempi kuin suomenkielinen muina hetkinä kuin iltapäivällä kello 13 - 15 välisenä aikana. Lisäksi englanti on selkeästi suomea suositumpi julkaisukieli myöhäisillasta. Englannin suurempi suosio iltaisin voi viestiä kielen olevan suomea vahvemmin kytköksissä vapaa-aikaan. Kuvasta näkyvä kolmipiikkinen julkaisuaktiivisuuden rakenne on tuttu koko Instagram-aineiston julkaisurytmiikkaa havainnollistavasta kuvasta 9.

Näiden tulosten perusteella näkee Helsingin Instagram-julkaisujen olevan vahvasti sidottu rakennettuun ympäristöön ja siten myös digitaaliseen kaupunkitilaan. Palvelualueiden yleisyys vahvistaa entisestään

Instagram-aineiston kytkeytymistä vapaa-aikaan. Myös kaupunkiluonto on edustettuna maankäyttöluokissa erityisesti meren, havu- ja sekametsän osalta, joten julkaisut kattavat fyysistä kaupunkitilaa metsineen ja merenrantoineen varsin hyvin. Maankäyttöluokkatarkastelun perusteella Instagram-aineistolla vaikuttaa olevan vahva side urbaaneihin maankäyttöluokkiin, joka antaa sille varsin hyvän mahdollisuuden kuvata digitaalista kaupunkitilaa, vaikkakin vain Instagram-alustalta nähden. Suomenkielisissä julkaisuissa näkyy lisäksi kytkös arkiseen työelämään maankäyttömuotojen kautta, jota puolestaan englanninkielisissä julkaisuissa ei yhtä selkeästi näy.

3.2 Lemmatisoinnin tulokset

Lemmatisointi onnistui ongelmitta, pois lukien aineistoihin jääneet yksittäiset vieraskieliset sanat, joita käytetyt lemmatisointimenetelmät eivät osanneet lemmatisoida, sekä muutaman sanan puhekieliset kirjoitusasut. Suomenkielisiä lemmatisoitavia kuvatekstejä oli 180 032 kappaletta ja ne koostuivat 36 733 uniikista sanasta. Vastaavasti englanninkielisiä lemmatisoituja kuvatekstejä oli 162 189 kappaletta ja ne koostuivat 21 197 uniikista sanasta. Kaikkien sanojen lemmatisoinnin lisäksi lemmatisoinnin tuloksista irrotettiin erillisiksi tekstiaineistoiksi lemmatisoidut adjektiivit sekä substantiivit ja verbit. Näin ollen lemmatisoiduista kuvateksteistä on kolme versiota: kaikki lemmatisoidut sanat, lemmatisoidut substantiivit ja verbit sekä kolmantena lemmatisoidut adjektiivit. Tämä tehtiin ensimmäisen aihemallinnuksen jälkeen, kun huomattiin, että aiheiden tulkintaa voisi helpottaa pelkästään substantiiveihin ja verbeihin keskittyvä aihemallinnus sekä kun päätettiin kokeilla, mikäli pelkästään adjektiiveihin keskittyvä aihemallinnus voisi toimia eräänlaisena “sävyanalyysina”. Substantiivi-verbi aihemallinnus ja adjektiivi aihemallinnus toteutetaan vain valituille kaupunginosille eikä koko Helsingin aineistolle. Lemmatisoinnin jälkeen aineistoista luotiin kaksi sanapilveä (kuvat 33 ja 34), joissa sanan yleisyyttä visualisoidaan kirjasinkoolla: mitä yleisempi sana aineistossa on, sitä suuremmalla kirjasinkoolla se kirjoitetaan sanapilveen. Sanapilvien sanastosta on poistettu hukkasanat. Sanapilvet ovat hyödyllinen työkalu, jolla saadaan eräänlainen ennakkokatsaus tekstuaalisen aineiston sisältöön yleisimpien sanojen kautta. Sanapilvissä näkyy muutama erikielinen sana, joka johtaa juurensa automaattisen kielentunnistuksen toimintalogiikasta. Suomenkielisestä aineistosta olisi voinut poistaa myös englanninkieliset hukkasanat.

Kuva 33. Sanapilvi 200 yleisimmästä sanasta suomenkielisessä Instagram-aineistossa. Yleisimmät sanat antavat pehmeän käsityksen niistä aiheista, mitä aihemallinnus tuottanee.

Kuva 34. Englanninkielisen aineiston lemmatisoiduista kuvateksteistä rakennettu sanapilvi. Mitä yleisemmin sana esiintyy aineistossa, sitä suuremmalla fontilla se on kirjoitettu.

3.3 Aihemallinnuksen tulokset

Aihemallin parametrit valittiin kokeilemalla useita eri iteraatio- ja läpikäyntiarvoja, kunnes hyvät arvot löytyivät. Tämän jälkeen samoja arvoja kokeiltiin pelkästään lemmatisoituja verbejä ja substantiiveja sisältävälle korpukselle, mutta koherenssipisteytys putosi huomattavasti, jonka vuoksi sen käyttö aihemallinnuksessa hylättiin koko Helsingin kattavalle aineistolle. Kohdealuekohtaisessa mallinnuksessa läpikäyntien määrää lisättiin ja aiheiden määrää tiputettiin, sillä kohdealueilla oli huomattavasti pienemmät julkaisumäärät kuin koko Helsingissä, mutta mallinnus toteutettiin lemmatisoitujen tekstien lisäksi niistä irrotetuille adjektiiveille kevyenä sävyanalyysinä ja substantiiveille sekä verbeille aktiviteettianalyysinä.

Mallien onnistumista mitattiin luomalla aihemalliin perustuva koherenssimalli, joka kertoo kuinka koherentteja (tai ymmärrettäviä) luodut aiheet ovat ihmiselle (Deshpande 2018). Koherenssipisteiden arvot laskettiin C_V-pisteytyksellä, joka on nollan ja yhden välillä esiintyvä arvo, jossa koherentimpi aihe tai malli saa läheltä arvoa yksi olevan pisteytyksen. Aiheiden lukumäärän nostaminen nostaa myös koherenssipisteitä huomattavasti, mutta tällöin aiheiden erottaminen semanttisesti toisistaan vaikeutuu, joka on erityisesti tämänkaltaisessa aineistossa ongelmallista, koska kuvatekstit ovat sanastoltaan ja sävyltään samankaltaisia (kuvat 33 ja 34).

Näiden tulosten valossa kaikki lemmatisoidut sanat sisältävä aineisto tarjosi parhaan alustan Helsingin alueelle muodostettaville malleille, sekä seuraavat parametrit osoittautuivat parhaiksi Instagram-aineiston mallintamiselle: 10 aihetta, 1400 iteraatiota ja 30 läpikäyntiä. Kohdealuekohtaiset mallinnukset toteutettiin 5 aiheella, 1400 iteraatiolla ja 60 läpikäynnillä. Samoja parametreja käytettiin molemmille kieliaineistoille. Aihemalli ei osaa luokitella pientä osaa julkaisuja mihinkään tiettyyn aiheeseen, jonka vuoksi se antaa julkaisulle yhtä suuren mahdollisuuden kuulua kaikkiin aiheisiin, esimerkiksi kymmenen aiheen aihemallissa tällainen julkaisu saisi 10 % todennäköisyyden kuulua kaikkiin aiheisiin. Jostain syystä nämä aiheluokitukseltaan epävarmat julkaisut luokittuvat ensimmäisen aiheen alle. Tämän vuoksi aihemallin tuloksia suodatettiin siten, että kymmenen aiheen mallissa hyväksyttiin vain yli 10 % varmuudella

onnistuneet ja viiden aiheen alueellisissa malleissa vain yli 20 % varmuudella onnistuneet luokitukset. Tämä suodatti lähes jokaisessa aihemallissa hieman yli tuhat julkaisua pois aiheen epäselvyyden vuoksi.

Aiheiden mallintamisen jälkeen tapahtuva nimeäminen on pitkälti tutkijan itsensä käsissä, jolloin se on varsin subjektiivinen prosessi. Aiheiden nimeämisen on todettu olevan hankalaa ja subjektiivista myös aikaisemmissa tutkimuksissa (Zhao et al. 2011; Lansley & Longley 2016). Aiheiden nimeämisessä apuna voi käyttää tilastollisia tarkasteluja aihefrekvensseistä ajassa sekä karttatarkasteluja, joissa katsotaan ovatko esimerkiksi uiminen-aiheen alle luokitellut julkaisut sijoittuneet lähelle merta, järviä, rantoja ja uimahalleja. Suoranaista tilastollista menetelmää aiheiden nimeämisen onnistumiseen ei ole ja sellaisen käyttäminen voisi olla ongelmallista, koska nimeäminen on pitkälti subjektiivinen prosessi, jonka myötä aiheisiin kytkettävät taustamuuttujat voivat muuttua tulkinnan muuttuessa. Sanojen moniselitteisyys ja polysemia voivat vaikeuttaa omalta osaltaan aiheiden nimeämistä, koska sanoilla ja lauseilla voi kontekstista riippuen olla useita merkityksiä, joiden havaitseminen vaikeutuu lemmatisoinnin ja aihemallintamisessa muodostuvien aiheelle tärkeiden sanojen listauksen myötä. Näistä ongelmista johtuen aiheiden nimeämisessä ja nimeämisen onnistumisessa on hyödynnetty yllä kuvatun kaltaisia frekvensseihin ja alueellisiin tarkasteluihin liittyviä keinoja. Esimerkiksi aiheen nimeämistä jouluksi voi tukea kyseisen aiheen kuukausittaisen esiintymisfrekvenssin tarkastelu, jossa aiheen frekvenssissä on selkeä julkaisupiikki joulukuussa. Vastaavasti urheilu-aiheen onnistumista voi tukea esimerkiksi se, että aiheen julkaisut ovat lähellä tunnettuja urheilupaikkoja ja esimerkiksi urheilutapahtumien aikoina.

3.3.1 Helsingin alueen tarkastelu

Molemmista aineistoista mallinnettiin 10 aihetta, jotka ovat eriteltynä alempana taulukoissa 10 ja 11. Suomenkielisellä aineistolla aiheiden mallintaminen onnistui hieman paremmin, sillä koherenssiarvot ovat korkeammat kuin englanninkielisellä aineistolla. Tämä todennäköisesti kertoo siitä, että suomeksi tehdyt julkaisut ovat kohdistetumpia aiheiltaan koko Helsingin mittakaavassa ja todellisten aiheiden määrä suomenkielisessä aineistossa on pienempi kuin englanninkielisessä aineistossa, sillä mallinnettavien aiheiden lukumäärää nostamalla voi parantaa aihekohtaisia koherenssipisteitä. Tämä koherenssipistetulos on hieman yllättävä, sillä aihemallinnukselle syötetty suomenkielinen aineisto oli noin 20 000 julkaisua

suurempi kuin englanninkielinen aineisto ja molemmat mallit luotiin samoin parametrein. Voisi olettaa, että lukumäärältään suuremmassa aineistossa olisi suurempi aiheiden kirjo pienempään aineistoon verrattuna. Tätä tulosta voi selittää jo mainittu seikka siitä, että suomenkielisissä julkaisuissa esiintyvät aiheet saattavat olla lukumäärältään rajatumpia jostain syystä. Toisena selityksenä voisi olla suomenkielisessä aineistossa esiintyvä sanastollisesti vähemmän kirjava kielenkäyttö, jonka myötä muodostuvat aiheet saavat paremman koherenssipisteytyksen, mutta näin todennäköisesti ei ole sillä suomenkielisessä aineistossa oli selkeästi enemmän uniikkeja sanoja englanninkieliseen aineistoon verrattuna. Kuten ylempänä mainittiin, suomenkielisessä aineistossa yksittäinen sana esiintyi keskimäärin 25 kertaa kun taas englanninkielisessä aineistossa yksittäinen sana esiintyi keskimäärin 35 kertaa.

Taulukko 10. Suomenkielisen aihemallinnuksen tulokset taulukossa. Taulukossa on aihekohtaisesti kymmenen tärkeintä sanaa tärkeysjärjestyksessä ylhäältä alas. Sanojen jälkeen on aiheesta tehtyjen julkaisujen määrä, kyseisen aiheen koherenssipisteytys (C_V) sekä aihetta kuvaava termi.

Aihe 1	**Aihe 2**	**Aihe 3**	**Aihe 4**	**Aihe 5**	**Aihe 6**	**Aihe 7**	**Aihe 8**	**Aihe 9**	**Aihe 10**
tehdä	joulu	päivä	ilta	käydä	paras	kesä	mennä	aurinko	pikkujoulu
työ	ihana	käynti	aamu	kiva	ruoka	helsinki	lähteä	syödä	kuva
saada	alkaa	with	treeni	löytyä	seura	suomi	pitää	valo	voida
klo	viikonloppu	kuvaus	vihdoin	festari	ystävä	eka	ottaa	lounas	itsenäisyys päivä
kaunis	kahvi	the	paikka	matka	rakas	keikka	ostaa	lahja	vähä
voida	koti	juttu	mieli	messut	kohta	hieno	jäädä	huomen	kylmä
päästä	tyttö	huikea	fiilis	kunnia	lapsi	syksy	kuva	paistaa	loma
ilo	nainen	alkaa	peli	nähdä	terveiset	joulukuu	kaa	talo	osata
tervetuloa	puhua	duuni	mahtava	perjantai	konsertti	sup	kiittää	tarina	kausi
mies	tähti	avaus	juhla	maanantai	sunnuntai	linna	saada	ihana	kauppa
16727	17677	17157	19425	17440	18448	17669	18817	16742	15676
0.38823	0.23238	0.28762	0.31775	0.33812	0.26586	0.32497	***0.42434***	0.24282	***0.21360***
Työ	**Joulu**	**Aloittaminen**	**Urheilu**	**Vierailu**	**Sosiaalinen kanssakäynti**	**Kesä**	**Arkielämä**	**Ruokailu**	**Juhlapyhä**

Taulukko 11. Englanninkielisten julkaisujen aihemallinnuksen tulokset taulukossa. Taulukossa on aihekohtaisesti kymmenen tärkeintä sanaa tärkeysjärjestyksessä ylhäältä alas. Sanojen jälkeen on aiheesta tehtyjen julkaisujen määrä, kyseisen aiheen koherenssipisteytys (C_V) sekä aihetta kuvaava termi.

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5	Aihe 6	Aihe 7	Aihe 8	Aihe 9	Aihe 10
nice	helsinki	night	year	finnish	start	finally	good	day	time
sun	light	work	love	fun	great	winter	thing	morning	christmas
walk	evening	lovely	happy	team	party	girl	friend	good	ready
lunch	coffee	sunday	place	game	people	wait	guy	finland	summer
beautiful	city	friday	helsinki	snow	office	find	house	photo	week
weekend	design	saturday	food	art	today	sea	birthday	feel	life
enjoy	will	well	yesterday	play	work	drink	live	helsinki	december
dinner	church	white	meet	experience	season	autumn	happen	today	awesome
ice	school	christmas	picture	lot	book	man	stuff	perfect	long
wine	tomorrow	sky	visit	rain	hard	delicious	music	cool	training
16302	14958	14325	17620	13712	14447	14014	14787	19899	14590
0.26224	0.25674	0.22171	***0.35458***	0.2989	0.33889	***0.21307***	0.34801	0.27047	0.26959
Ruokailu	**Helsinki**	**Viikonloppuu**	**Uusivuosi**	**Suomalaisuus**	**Juhlinta**	**Talvi**	**Ystävät**	**Aamu**	**Joulu**

Taulukoista 10 ja 11 on nähtävissä koko Helsingin alueen Instagram-julkaisujen kuvatekstien aihemallinnuksen tulokset. Aiheiden numeroiden alla on kymmenen aihekohtaisesti tärkeintä sanaa tärkeysjärjestyksessä ylhäältä alas. Sanojen jälkeen on julkaisujen määrä, jotka kuuluvat kyseiselle aiheelle. Julkaisumäärien alla on aiheen onnistumista kuvaava koherenssipisteytys esitettynä C_V-arvona välillä 0-1, joissa lähellä nollaa oleva pisteytys kertoo aiheen koherenssin olevan heikkoa ja lähellä yhtä oleva pisteytys kertoo aiheen koherenssin olevan vahvaa. Minkään aiheen koherenssi ei ole kovin suurta, joka ei yllätä, sillä kymmenen aiheen mallinnus kattaa koko Helsingin alueen ja koko aineiston aikavälin, jolloin muodostuvat kymmenen aihetta ovat väkisinkin varsin yleispiirteisiä molemmilla kielillä. Parhaiten onnistuneet aiheet ovat arkielämä suomen- ja uusivuosi englanninkielisessä aihemallissa. Huonoimmin onnistuneet mallit ovat juhlapyhä suomen- ja talvi englanninkielisessä aihemallissa. Aiheiden lukumäärän nostaminen myös nostaisi koherensseja, sillä tällöin yleispiirteisten aiheiden tilalle muodostuisi tarkempia aiheita, mutta aiheiden lukumäärän noustessa niiden semanttinen erottelu ja nimeäminen vaikeutuvat. Myös alueellinen ja/tai ajallinen rajaaminen nostaisi aiheiden koherensseja. Alueelliset aihemallit on kappaleessa 3.3.2, ajassa rajattuja aihemalleja ei tässä työssä tehty.

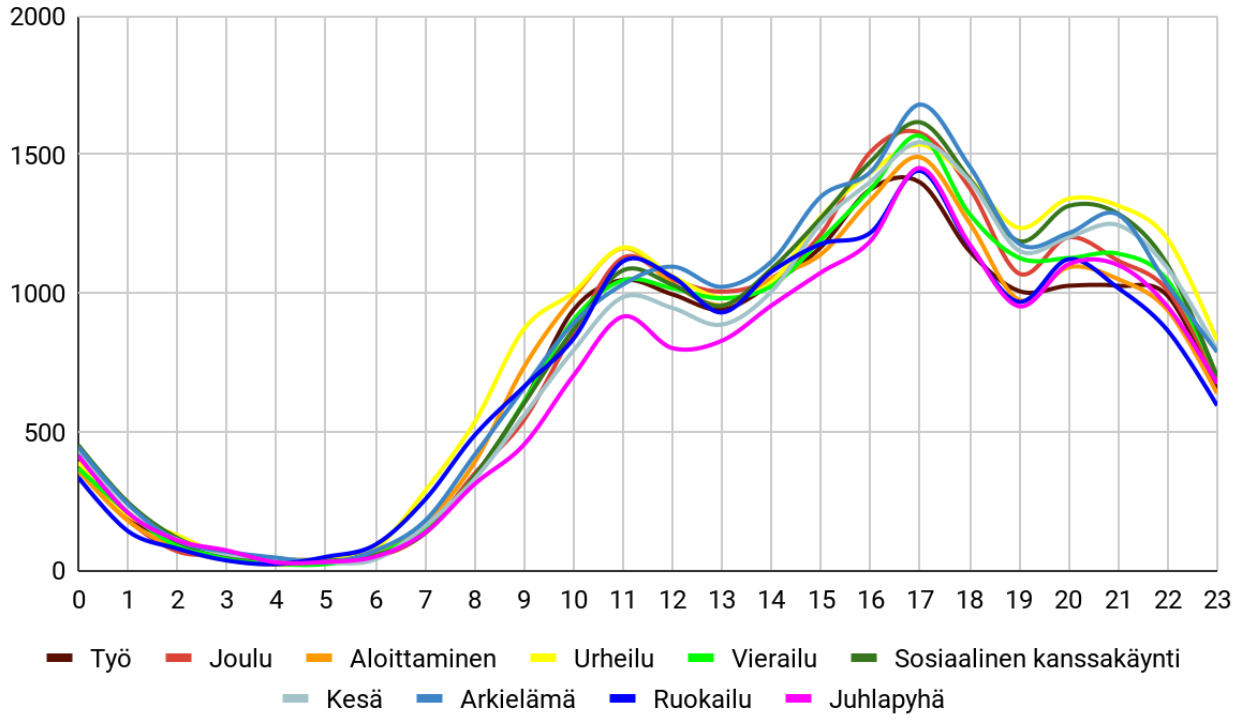
Aiheita tarkastellessa on selvää, että aiheet ovat samankaltaisia keskenään englannin- ja suomenkielisissä Instagram-julkaisuissa: ruokailu esiintyy molemmissa sekä sosiaalinen kanssakäynti, ystävät, vierailu ja

juhlinta ovat semanttisesti varsin lähekkäisiä aiheita. Ruokailu-aiheiden sanastot myös viittaavat ravintoloissa ja sosiaalisissa tilanteissa tapahtuvaan ruokailuun, jonka voisi nähdä normaalista arjesta poikkeavana ruokailutapahtumana. Myös selkeitä eroja löytyy: esimerkiksi vain suomenkielissä aiheissa on vahvasti arkeen liittyviä aiheita, kuten työ, arkielämä ja urheilu. Arkeen suoranaisesti liittyviä aiheita ei englanninkielisissä aiheissa ole. Suomenkielisissä aiheissa on vuodenajoista kesä, englanninkielisissä talvi. Englanninkielisissä aiheissa on Helsinki ja suomalaisuus omina aiheinaan, joita ei suomenkielisistä aiheista löydy. Suomeksi kirjoitettujen julkaisujen aiheissa on myös vahvasti kulttuurisidonnaisia juhlia mainittu juhlapyhä-aiheen alla kuten itsenäisyyspäivä ja pikkujoulut.

Aiheita kuvaavien termien määrittely on hankalaa, erityisesti niissä tapauksissa, joissa tärkeimmät sanat eivät vaikuta johdonmukaisilta toisiinsa nähden. Aiheiden määrittelyn hankaluus ei ole yllättävää, sillä koherenssipisteet ovat heikkoja. Toisaalta esimerkiksi taulukon 10 aihe 8 on onnistunein aihe koherenssipistein mitattuna, mutta sille merkittävimmät sanat tekevät aiheen nimeämisestä varsin hankalaa. Sanoja tarkastelemalla arkielämä lienee osuvin nimi aihe 8:lle. Taulukossa 11 vastaavanlainen vaikea nimeäminen on aiheella 4, sillä aiheelle merkittävimmät sanat tuovat mieleen yleisen hyvän uudenvuoden toivotuksen, mutta muut sanat liittyvät ruokaan, tapaamiseen ja Helsinkiin. Aiheen 4 kolmessa tärkeimmässä sanassa olivat sanat “happy” ja “year”, jotka tuovat mainitun yleisen uudenvuoden toivotuksen mieleen. On huomautettava, että “new”-sana on englanninkielinen hukkasana, joten sitä ei ole luettu mukaan aihemalliin eikä siten näy aiheissa. Aihe 4 voi olla onnistuneesti nimetty uusivuosi-aiheeksi, mikäli sillä on selkeä julkaisupiikki vuodenvaihteessa, joka sillä on (kuva 39). Tästä huolimatta se pitää sisällään varmasti myös uuteenvuoteen suoranaisesti liittymättömiä julkaisuja. Tämän kaltaisia ongelmatilanteita voisi vähentää jatkossa esimerkiksi liittämällä useasta erillisestä sanasta koostuvat käsitteet yhteen jollain merkillä. Aiheet voivat pitää sisällään myös toisilleen näennäisesti vastakohtaisia sanoja, esimerkiksi taulukon 10 viimeinen aihe, joulu, sisältää sanan summer. Tämä sanojen anakronistisuus ei kerro aiheen epäonnistumisesta, sillä esimerkiksi julkaisussa voidaan toivottaa hyvää joulua, mutta samalla toivoa kesän pikaista saapumista.

Sanojen epäselvien aiheyhteyksien, anakronistisuuden ja muiden mahdollisten vastakohtaisten sanojen vuoksi aiheiden nimeäminen on tämän työn aihemallien vaikeimpia ja subjektiivisimpia vaiheita, mutta oli odotettavissa teoriakirjallisuuden perusteella (Batty 2010; Graham & Zook 2011; Kellerman 2014, 2015; Malecki 2017; Rose 2017). Aihemallin tuloksena syntyneiden aiheiden nimeäminen on tutkijan "käsityöksi" jäävä tehtävä, jonka vuoksi aiheille annetut nimet ovat subjektiivisia ja siksi niihin tulee suhtautua varauksella sekä kriittisesti. Aiheiden nimeäminen tekee kuitenkin tuloksien tarkastelusta ymmärrettävämpää, vaikka se onkin subjektiivista ja periaatteessa valinnaista. Molemmissa aineistoissa esiintyviä aiheita ovat joulu, eri vuodenajat ja sosiaalinen kanssakäyminen. Englanninkielisessä aineistossa esiintyy mielenkiintoisesti aiheena suomalaisuuden kokeminen, joka lienee osa sitä eksotiikkaa, mitä täällä vierailevat turistit haluavat kokea, mutta voi myös viestiä suomalaisten "kotipaikkaylpeydestä", jota he viestivät kansainvälisesti laajasti käytetyllä kielellä, englannilla.

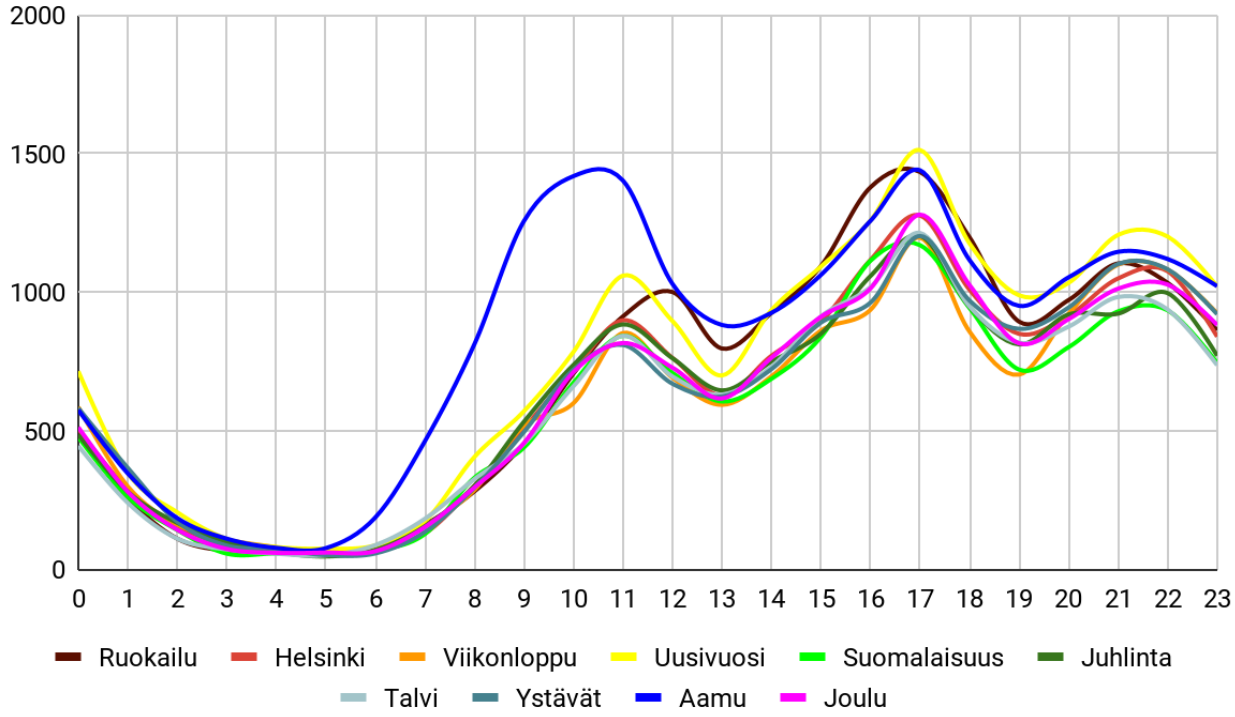
3.3.1.1 Aiheet ajassa



Kuva 35. Suomenkieliset aiheet tunneittain, josta on nähtävissä yleiset ruokailuajat selkeästi, joka on myös nähtävissä kuvassa 9 ylempänä. Aiheiden esiintymistiheydet seuraavat melko tiiviisti toisiaan ja oikeastaan vain urheilu- ja juhlapyhäaiheet erottautuvat muista aiheista silminnähtävästi.

Suomen- ja englanninkielisten aiheiden ajallista jakaumaa (kuvat 35 ja 36) tarkastellessa kellonajan mukaan paljastuu samoja piirteitä, joita paljastui aikaisemmin koko käsittelemätöntä aineistoa visualisoidessa

kellonajan mukaan kuvassa 9. Myös aiheiden kellonaikatarkasteluissa aamuyö on hiljaisinta julkaisuaikaa ja kolme julkaisupiikkiä muodostuu yleisten ruokailuaikojen kohdille. Suomenkielisten aiheiden vuorokausirytmistä paljastuu arjen läsnäolo seuratessa aiheiden suosion hierarkiaa esimerkiksi aamupäivät alkavat urheiluaiheella, keskipäivä täyttyy arkielämästä ja illalla urheilu sekä sosiaalinen kanssakäynti nousevat arkielämää yleisemmiksi aiheiksi Instagramissa. Arkikytköstä vahvistaa työaihe jo itsessään, mutta sen ajallinen luonne vaikuttaa varsin loogiselta, sillä työpäivän päätteeksi aiheen yleisyys putoaa eikä muodosta kolmatta piikkiä muiden aiheiden tavoin. Kellonaikaan perustuva tarkastelu on muutaman aiheen kuten joulun, kesän ja juhlapyhän osalta hieman omituista, mutta kuukausiin perustuvassa visualisoinnissa (kuva 39) kyseisten aiheiden tarkastelu ja nimet tuntuvat huomattavasti järkevämmiltä.

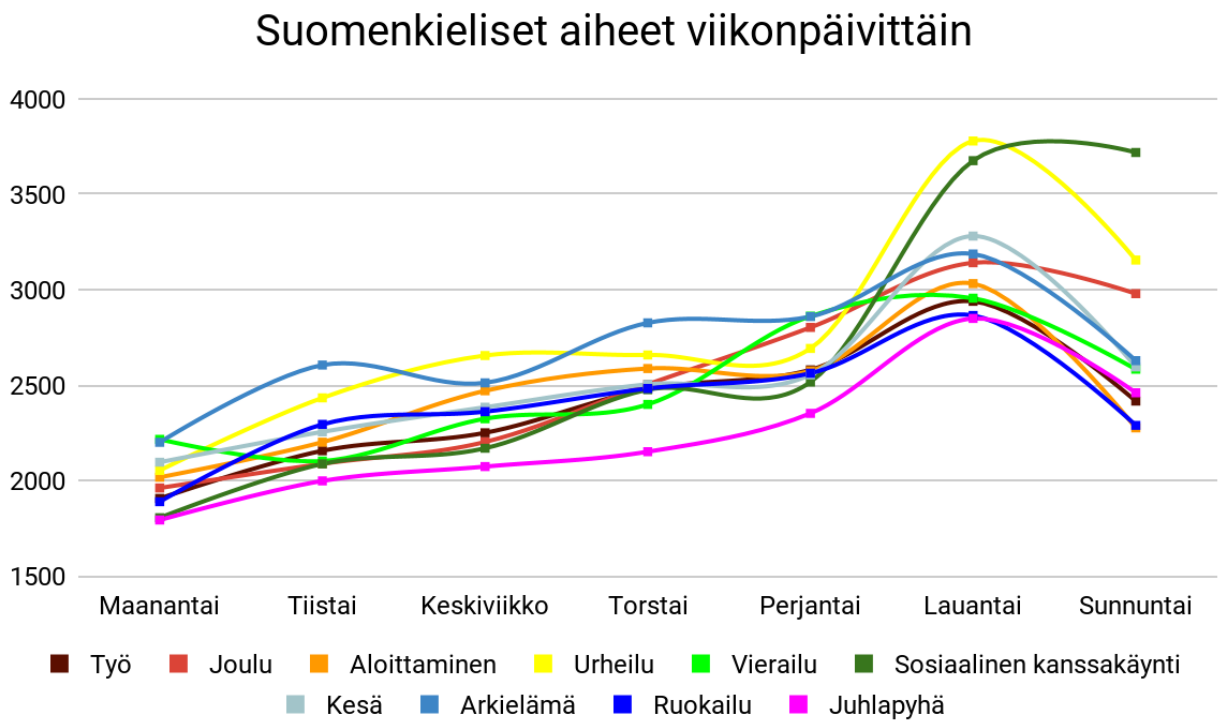


Kuva 36. Englanninkieliset aiheet tunneittain, josta myös on nähtävissä selkeästi yleiset ruokailuajat. Suomenkielisestä aineistosta poiketen kolme aihetta erottuu kellonajallisesti yleisimpinä aiheina.

Englanninkielinen aineisto pitää sisällään selkeämmin massasta erottuvia aiheita kuten ruokailu, uusivuosi ja aamu. Aamu-aihe on nimetty melko osuvasti, sillä se on aamuisin selkeästi suurin aihe, mutta kello 11 eteenpäin aamu alkaa menettää suosiotaan, mutta pysyy suosituimpien aiheiden joukossa lävitse koko vuorokausitarkastelun. Aamu-aiheen lähes yhtä korkeat piikit julkaisumäärissä aamu- ja iltapäivällä viestii aiheen yleispiirteisyydestä ja osittaisesta nimeämisen epäonnistumisesta. Kyseiselle aiheelle morning -sana

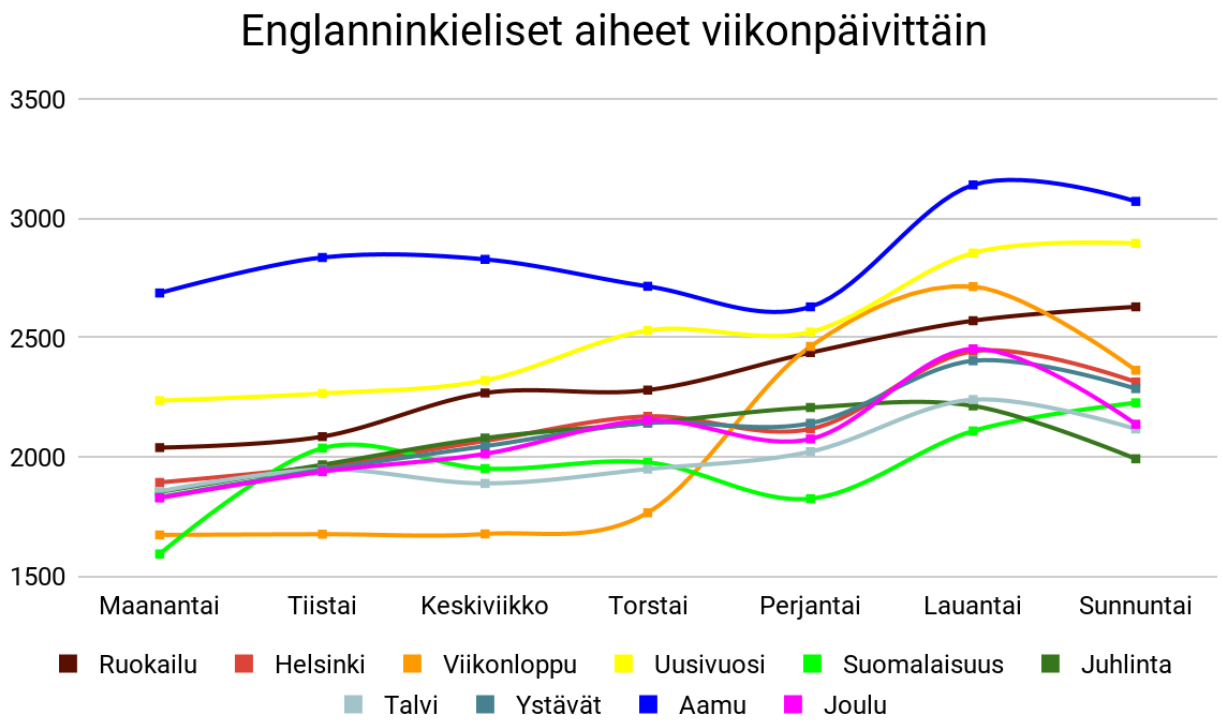
on yksi tärkeimpiä sanoja, mutta aiheen muut sanat eivät liity suoraan nimenomaan aamuun vaan lukuisiin muihin asioihin. Myös uusivuosi-aihe on yleisimpiä aiheita vuorokausitarkastelussa, mutta senkin osalta nimeämisen onnistuminen on hieman kyseenalaista. Uusivuosi-aihe pitää sisällään lukuisia sanoja, jotka eivät välttämättä liity uuteen vuoteen, jonka vuoksi julkaisut jotka eivät liity uuteenvuoteen, mutta sisältävät muita aiheelle tärkeitä sanoja luokittuvat aiheen alle. Tästä nimeämisen hankaluudesta huolimatta uusivuosi-nimi on aihetta melko hyvin kuvaava, kun huomioi yllämainitun useasta sanasta muodostuvien käsitteiden käsittelyn vaikeuden ja kuvan 40 kuukausitarkastelun perusteella aiheelle annettu nimi on silti sitä melko hyvin kuvaava. Ruokailu-aihe on myös muista englanninkielisistä aiheista yleisyydellään erottuva aihe ja se nouseekin toiseksi vahvimmaksi kello 12 ja vahvimmaksi kello 16, jotka eivät ole yllättäviä aikoja, sillä näihin aikoihin usein ruokaillaan. Uusivuosi-aihe nousee yleisimmäksi aiheeksi tasan kello 17 ja kello 21-22 välisinä aikoina. Myös englanninkielisessä aineistossa suurin piikki on kello 17 aikaan ja myöhäisillan piikki on pääpiirteittäin suurempi kuin aamun piikki kaikilla muilla aiheilla paitsi aamu-aiheella. Englanninkielisistä aiheista on vaikea erottaa yhtä selkeää pienintä aihetta kellonajan mukaan tosin suomalaisuus ja viikonloppu ovat useaan otteeseen pienimpinä.

Tunneittain molempia aineistoja toisiinsa vertaillessa huomaa, että aineistoissa esiintyy samankaltainen rakenne julkaisupiikkien muodossa, mutta englanninkielisessä aineistossa on selkeästi vahvemmin erottuvia yksittäisiä aiheita. Toisin sanoen suomenkielinen aineisto on aiheiden määrien suhteen yhdenmukaisempi ja tasaisempi kuin englanninkielinen aineisto. Suomenkielisessä aineistossa näkyi myös arkiseen elämään kytköksissä olevien aiheiden yleisyyden ja niiden rytmiikan noudattavan yleisiä vuorokausirytmejä: urheilua harrastetaan usein aamuisin ja iltaisin, arkielämä tapahtuu päivällä ja työ-aiheen putoaminen kello 17 jälkeen vastaa pitkälti yleisesti käytössä olevan 9 - 17 työajan päättymistä. Englanninkielisessä aineistossa oli hieman enemmän yleispiirteisiä aiheita, jonka vuoksi muutaman aiheen, uudenvuoden ja aamun, nimi kuvaa aihetta vain sumeasti. Molemmissa aineistoissa oli yksi aamuisin yleinen aihe, joka pitkälti säilytti korkean yleisyytensä läpi eri kellonaikojen. Ylipäätään temporaalinen rytmiikka paljastaa Helsingin pulssin (ks. myös Batty 2010), joka on suomeksi ja englanniksi kirjoitetuissa julkaisuissa samankaltainen kuin kaikkien julkaisujen pulssi (kuva 9).



Kuva 37. Suomenkielisten aiheiden esiintymistiheys viikonpäivittäin esitettynä. Kaikkien aiheiden esiintymistiheys nousee viikonloppua kohti, mutta useassa tapahtuu notkahduksia. Viikonlopusta lauantai on selvästi vilkkain päivä kaikkien aiheiden osalta.

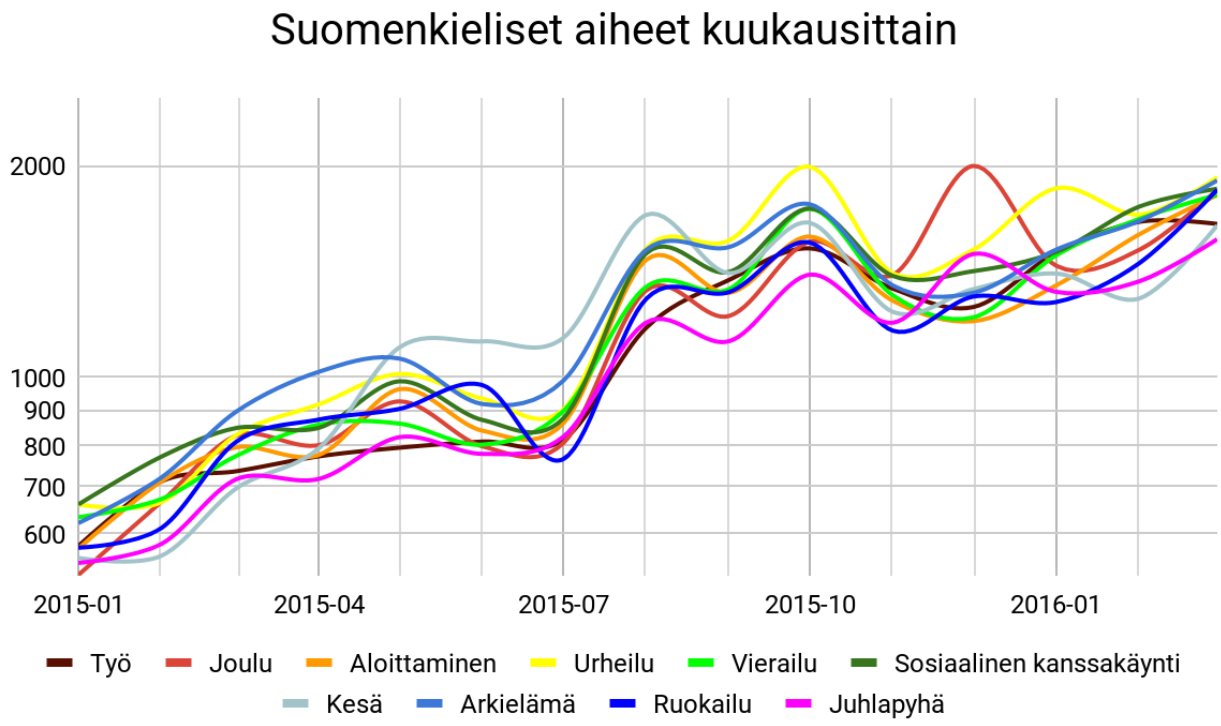
Aiheiden ajallinen jakauma viikonpäivittäin (kuvat 37 ja 38) paljastaa samankaltaisen rakenteen kuin tunneittain tarkasteltuna: suomenkielinen aineisto on vaihteluiltaan tasaisempaa ja englanninkielisessä aineistossa on selkeästi muista aiheista erottuvia aiheita. Molemmissa tarkasteluissa on nähtävissä kaupungin pulssi (Batty 2010), jonka korkein kohta on viikonloppuisin. Suomenkielisistä aiheista sosiaalinen kanssakäynti on ainoa, joka kasvaa jatkuvasti läpi viikon ja myös viikonloppuna lauantaista sunnuntaihin. Yleisimpinä aiheina suomenkielisessä aineistossa vuorottelevat urheilu ja arkielämä koko viikon, kunnes arkielämä tippuu viikonlopun tullessa ja sosiaalinen kanssakäynti nousee urheilun rinnalle lauantaina, mutta jatkaa kasvuaan tullen suvereenisti yleisimmäksi aiheeksi sunnuntaisin.



Kuva 38. Englanninkielisten aiheiden viikonpäivittäinen esiintymistiheys. Kolme päivittäin yleisintä aihetta pitävät asemansa miltei läpi viikon, mutta viikonloppu-aihe nousee erittäin selkeästi perjantaisin ja on yleisimpiä aiheita lauantaisin ja sunnuntaisin. Lauantai on kaikkien aiheiden osalta vilkkain viikonpäivä.

Englanniksi kirjoitettujen julkaisujen aiheissa (kuva 38) ei ole samaa aiheiden välistä yhtäläisyyttä viikonpäivittäin kuin suomeksi kirjoitetuissa aiheissa. Hieman omituisesti osa aiheista lopettaa määrällisen kasvunsa tai putoaa äkisti perjantaisin, jonka jälkeen ne nousevat jyrkästi lauantaina. Tämän taustalla voi olla viikonloppu-aiheen erittäin selkeä nousu perjantaisin. Lähestulkoon kaikki aiheet laskevat määrällisesti lauantaista sunnuntaihin, joka vastaa aikaisempia havaintoja käytössä olevasta Instagram-aineistoista (kuvat 8 ja 9), mutta uusivuosi, ruokailu ja suomalaisuus kasvavat myös lauantaista sunnuntaihin. Täysin nousujohteinen viikko aiheista on ruokailulla. Aiheista kolme yleisintä pitävät asemansa lähes koko viikon, mutta perjantaina suuren osan viikkoa pienimpänä aiheena ollut aihe, viikonloppu, nousee kolmanneksi suurimmaksi perjantaisin ja lauantaisin. Viikonloppu-aiheen suosio viikonloppuisin ei ole yllättävää, mutta se vahvistaa nimeämisen onnistumisen ja on siksi mielenkiintoinen havainto. Uusivuosi-aiheen nousu läpi viikon ei varsinaisesti eroa myös muilla aiheilla näkyvistä kasvutrendeistä viikonloppua kohden, mutta tämä viikonpäivittäinen tarkastelu asettaa aiheen nimeämisen onnistumisen varsin kyseenalaiseksi. Muiden aiheiden välillä ei tapahdu kovin suuria muutoksia, mutta suomalaisuus-aihe vaikuttaa olevan muihin

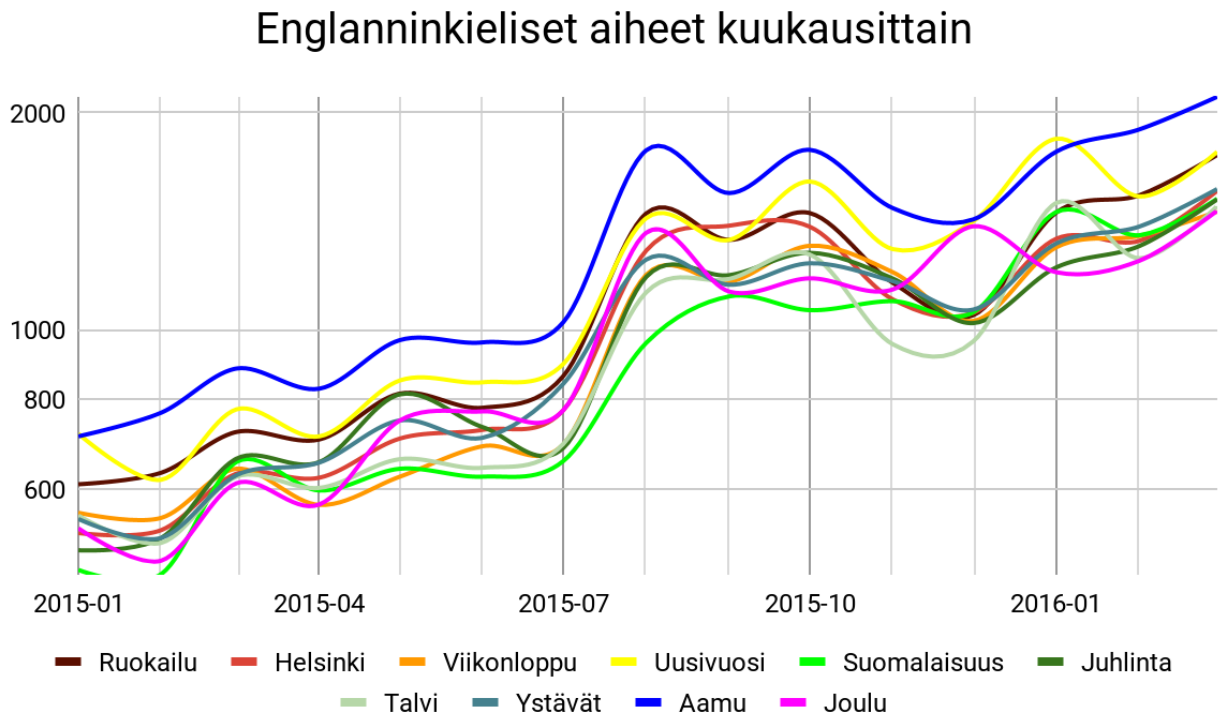
aiheisiin verrattuna erilainen, sillä se saavuttaa kaksi piikkiä viikon aikana: tiistaisin ja sunnuntaisin. Näille piikeille ei ole mitään ilmeistä syytä, mutta aiheen huomioiden se voi liittyä mahdollisiin matkailijoiden tekemiin julkaisuihin. Vapaa-ajan matkustuksessa viikon sisäinen tavallinen arjen ja viikonlopun välinen rytmitys ei näy yhtä voimakkaasti, joka voisi selittää tiistaisin tapahtuvaa piikkiä.



Kuva 39. Suomenkielisten julkaisujen määrä niiden aiheisiin jaoteltuna ja esitettynä kuukausittain. Aiheissa on kolme selkeää piikkiä: kesä-aihe elokuussa 2015, urheilu-aihe lokakuussa 2015 ja tammikuussa 2016 sekä joulu-aihe joulukuussa 2015.

Kun aiheiden ajallista kehitystä tarkastelee kuukausittain läpi koko aineiston ajallisen kattavuuden vuoden 2015 tammikuusta vuoden 2016 maaliskuuhun saakka (kuvat 39 ja 40), aineistojen välillä on nähtävissä samankaltaisia piirteitä julkaisumäärien muutoksissa mutta myös aikaisemmin huomatut piirteet näkyvät. Suomenkielisissä aiheissa yhtä hallitsevaa aihetta ei ole, tosin pisimpään yleisimpänä aiheena on kesä toukokuusta elokuuhun ja samalla se on talvisin pienimpiä aiheita, joka vahvistaa nimeämisen onnistumista. Tämä yleisimmän aiheen vaihtelu voi kertoa suomenkielisten julkaisujen olevan jossain määrin vuodenaikasidonnaisia. Kesän jälkeen urheilu on yhteensä kolme ja arkielämä kaksi kuukautta yleisimpinä aiheina. Urheilu-aihe käy läpi kaksi piikkiä, joista toinen on lokakuussa ja toinen tammikuussa. Tammikuisen piikin taustalla lienee yleinen tapa aloittaa urheilun harrastaminen uudenvuodenlupauksien tekemisen

myötä, lokakuisen piikin taustasyy ei ole yhtä ilmeinen. Joulu-aiheella on piikki joulukuussa, joka vahvistaa aiheen nimeämisen onnistumista. Useilla suomenkielisillä aiheilla on kaksi piikkiä: elo- ja lokakuussa 2015. Elokuinen piikki liittynee Instagramin siirtymisestä Facebookin omistukseen ja siitä alkunsa saaneeseen suureen käyttäjämäärän kasvuun (Cvetojevic et al. 2016). Tämä elokuinen kasvupyrähdys näkyy myös englanninkielisessä aineistossa. Lokakuisen piikin aiheuttajasta ei ole varmuutta, mutta Helsingissä järjestetään lokakuisin Viini- ja ruokamessut, joka voi selittää lokakuista piikkiä.



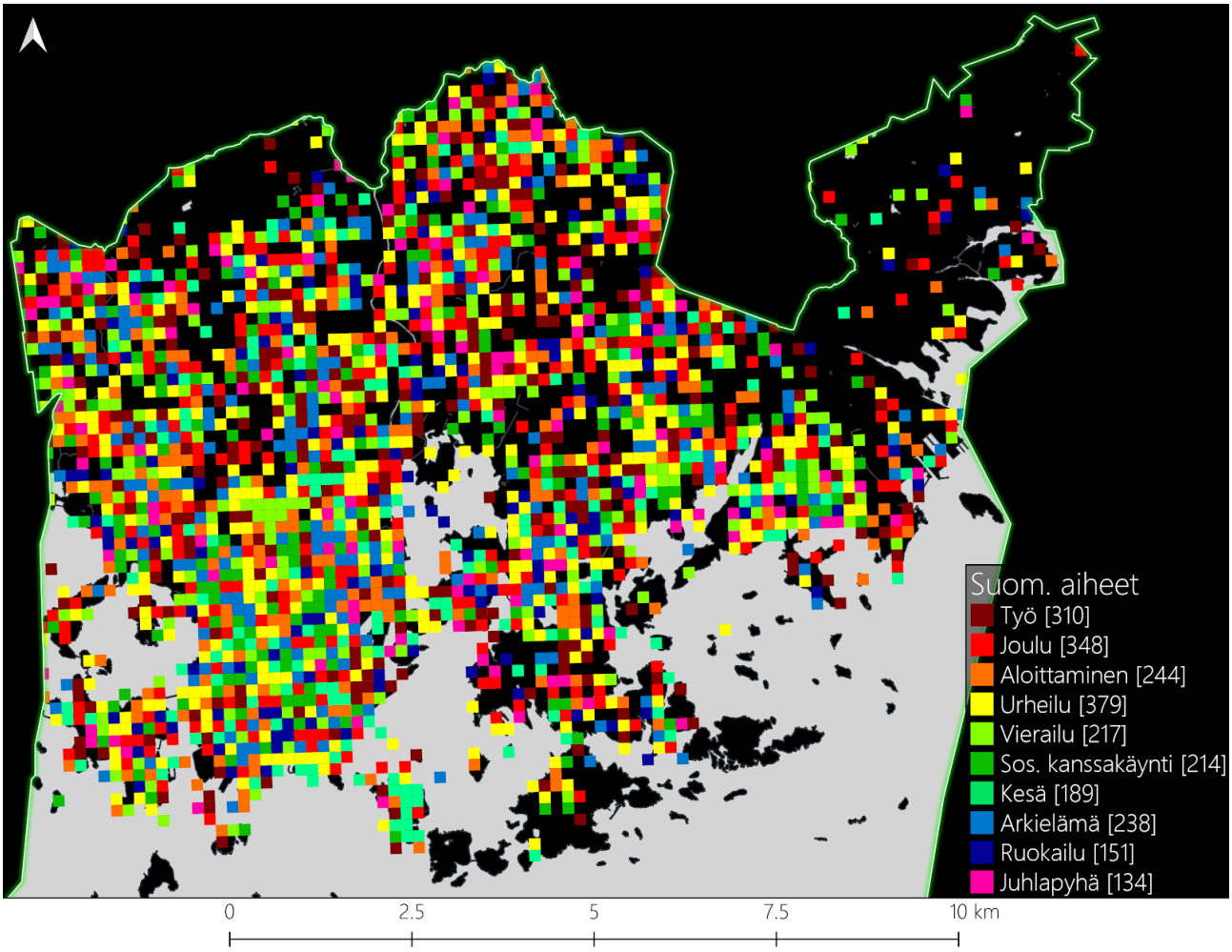
Kuva 40. Englanninkielisten julkaisujen määrä niiden aiheisiin jaoteltuna ja esitettynä kuukausittain. Kolme aihetta erottautuu kuvassa selkeästi muista suuremmilla määrillään: aamu, uusivuosi ja ruokailu.

Englanninkielisessä aineistossa on muutama aihe, jotka ovat selvästi yleisempiä läpi koko aineiston, kun taas suomenkielisessä aineistossa suurimman aiheen asema vaihtelee useaan otteeseen. Kenties englanninkielinen aineisto ei ole aiheiltaan yhtä vuodenaikasidonnainen kuin suomenkielinen aineisto. Toisaalta myös englanninkielisessä aineistossa näkyy vuodenaikoihin liittyviä aiheita kuten talvi joulu, joista joulu esiintyy myös suomenkielisessä aineistossa. Englanninkielisistä aiheista ylivoimaisesti yleisin on aamu, joka dominoi koko ajanjaksoa vuoden 2016 tammikuuta lukuun ottamatta. Mielenkiintoisena lisähuomiona aamu-aiheen etäisyys seuraavaksi suurimpaan aiheeseen vaikuttaa pysyvän samankaltaisena kuukaudesta toiseen. Aamu-aiheen lisäksi englanninkielisissä aiheissa aihemassasta erottuvat uusivuosi- ja

ruokailuaiheet. Uusivuosi-aihe muodostaa selkeän piikin tammikuussa 2016, mutta myös lokakuussa 2015. Tammikuinen piikki vahvistaa uusivuosi-aiheen muuten ongelmalliseksi todetun nimen nimeämisen onnistumista hieman. Ruokailu on toiseksi suosituin englanninkielinen aihe helmikuussa 2015 ja elokuussa 2015, mutta mitään ilmeistä syytä näille suosion hetkille ei ole. Myös englanninkielisen aineiston joulu-aihe muodostaa selkeän piikin vuoden 2015 jouluna, tosin suomenkielinen vastine muodostaa huomattavasti korkeamman piikin ollen kyseisen aikajakson ylivoimaisesti suosituin aihe. Joulun aikana englanninkielisessä aineistossa kaikki muut aiheet laskevat vuoden 2015 joulukuussa paitsi joulu-aihe, kun taas suomenkielisessä aineistossa joulun lisäksi juhlapyhä nostaa yleisyyttään selkeästi.

Ajallisesti tarkasteltuna aiheet, kuten julkaisuaktiivisuuskin, ovat painottuneet vapaa-ajalle ja viikonloppuun, joka kertoo kaupungin pulssin Instagram-aineiston kautta nähtynä olevan vahvasti vapaa-aikasidonnainen. Ajallinen tarkastelu toi aiheiden nimeämiseen lisää varmuutta muutaman aiheen kohdalla, kun taas muutaman muun aiheen osalta aikatarkastelu toimi samanaikaisesti nimeämistä tukevasti, että sitä vastaan. Ylipäätään aiheiden aikatarkastelu paljasti mielenkiintoisia piirteitä molempien kielien aineistosta. Vuorokausitarkastelussa julkaisuaktiivisuuden kolmipiikkinen temporaalisuus saa syvyyttä, kun eri aiheiden yleisyys muuttuu aamusta iltaan mentäessä. Suomenkielisestä aineistosta on nähtävissä sen kytkeytyminen arkielämään arkisten aiheiden kautta. Esimerkiksi työ ja arkielämä, jotka ovat Jan Gehlin luokittelun mukaisesti pakollisia aktiviteetteja, ovat yleisimpiä aiheita huolimatta siitä, mitä aikaväliä käyttää tarkasteluun. Englanninkielisessä aineistossa taas ei varsinaisesti ole pakollisia aktiviteetteja ole ja sen myötä aiheissa näkyy vahvemmin vapaaehtoisiin aktiviteetteihin nojaavat aiheet kuten ruokailu, juhlinta, uusivuosi, joulu ja ystävät.

3.3.1.2 Aiheet spatiaalisesti

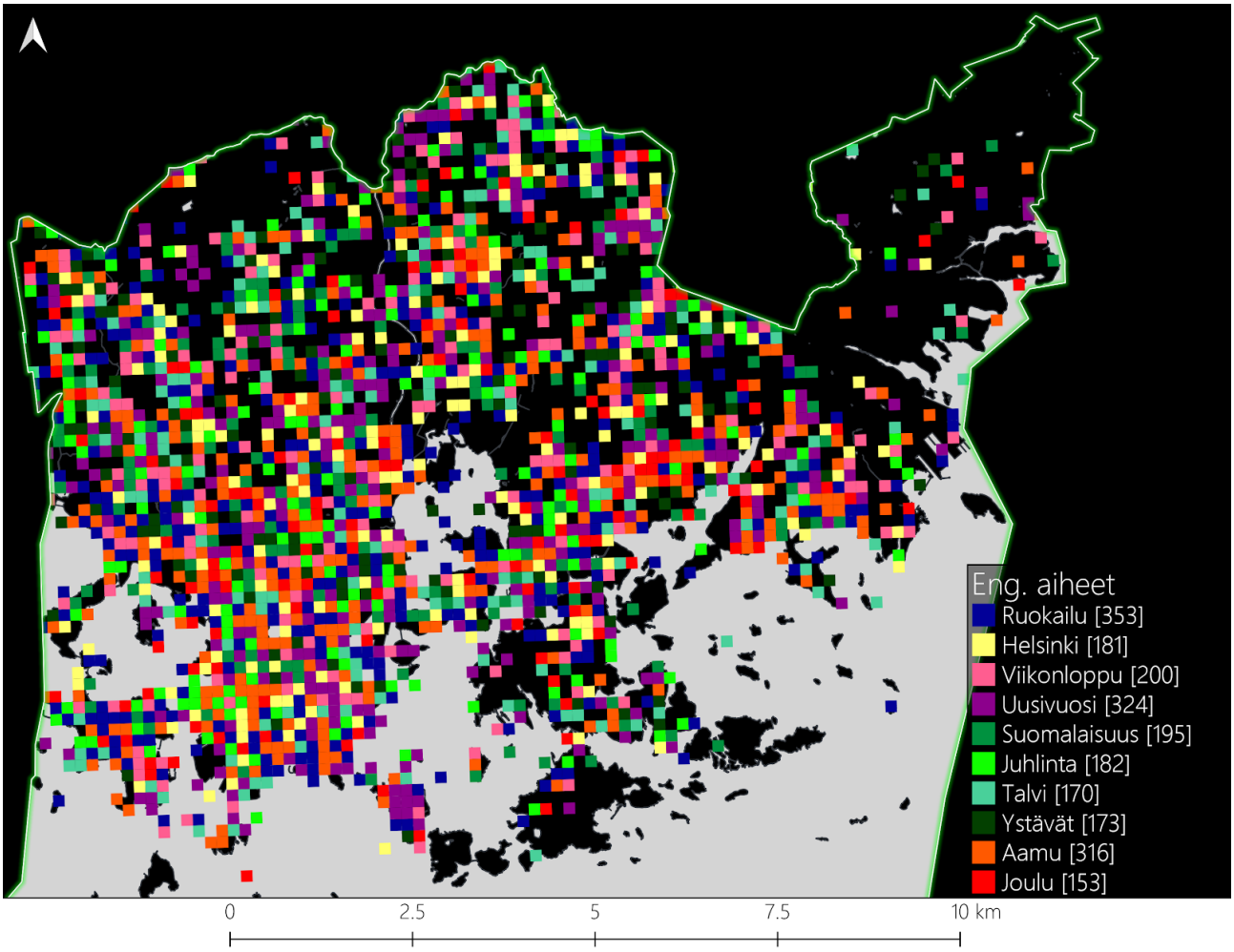


Kuva 41. Suomenkielisen aineiston aiheiden levinneisyys aggregoituna YKR-ruudukkoon. Aineistosta ei ole erotettavissa selkeää alueellista rakennetta, vaan rakenne näyttää päällepäin satunnaiselta. Hakasuluissa solujen lukumäärä.

Kuvan 41 karttaan on visualisoitu suomenkieliset mallinnetut 10 aihetta, jotka ovat aggregoitu 250 x 250 metrin YKR-ruudukkoon, siten että ruudussa oleva yleisin eli dominoiva aihe on merkitty eri värein. YKR-ruudukkoon visualisoituja aiheita on vaikea tulkita, sillä tällä tarkastelutasolla kuviointia ja klusterointia on melko vaikea havaita nopeasti, mutta maltillisella tarkastelulla spatiaalisia piirteitä pystyy havainnoimaan.

Suomenkielisistä aiheista selkeää klusteroitumista voi havaita esimerkiksi työ- ja vierailu-aiheilla. Työ-aihe klusteroituu jonkin verran Kallion alueella, Kampissa ja Lahdenväylän varressa keskustasta koilliseen. Vierailu-aihe vaikuttaa klusteroituvan Hartwall Arenan ympäristössä. Kesä-aihe on klusteroitunut erityisen vahvasti Suomenlinnan alueelle sekä Käpylän eteläpuolelle suurin piirtein Kumpulan siirtolapuutarhan alueelle. Lisäksi sosiaalisen kanssakäynnin aihe on keskittynyt keskusta-alueelle Kampin ja Kallion välille.

Tämän lisäksi urheilu-aiheella on muutamia mielenkiintoisia spatiaalisia piirteitä kuten Olympiastadionin ympäristö, Eiranrannasta Kaivopuistoon kulkeva suora Etelä-Helsingissä, Kumpulanlaaksossa sijaitseva klusteri sekä Lapinlahden alueella Kampin ja Lauttasaaren välissä. Kaikissa sijainneissa on suosittuja lenkkeilypolkuja ja Lapinlahden vieressä on Salmisaaren liikuntakeskus, sekä Kumpulanlaakson vieressä on UniSportin ylläpitämä liikuntakeskus. Aiheista arkielämä, ruokailu ja joulu eivät näytä olevan klusteroituneet juuri lainkaan vaan dispersoituneet ympäri Helsinkiä melko tasaisesti.



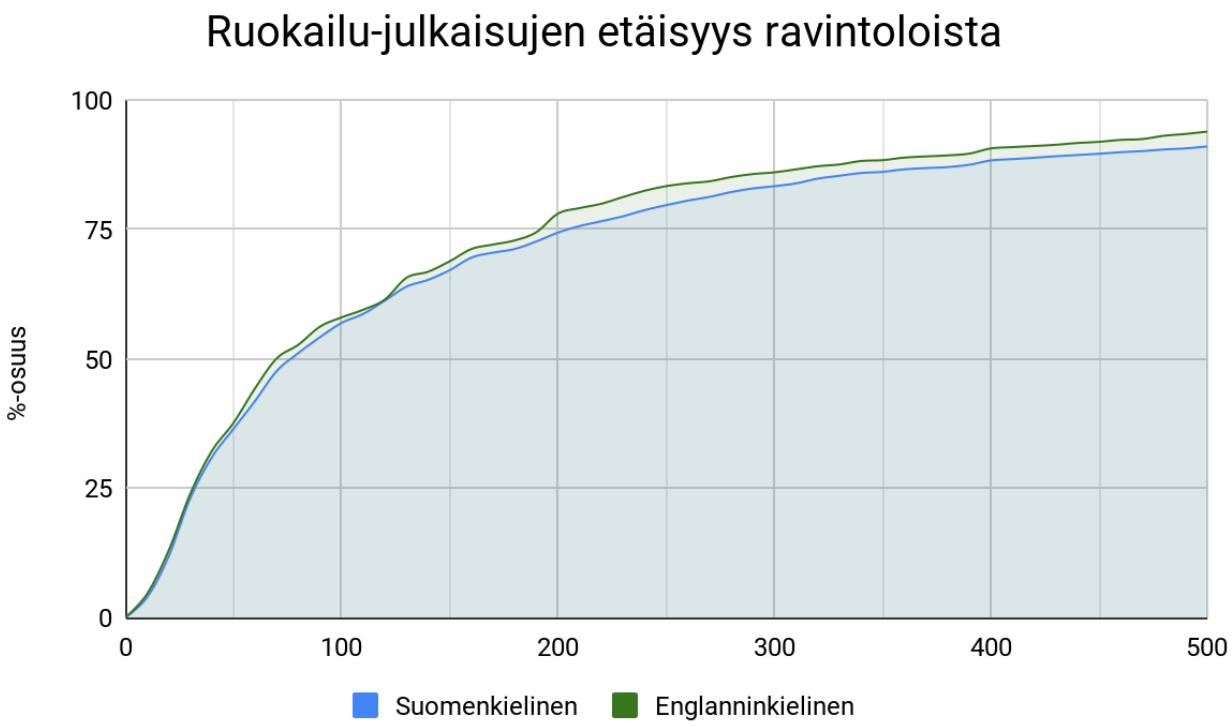
Kuva 42. Englanninkielisestä aineistosta mallinnetut 10 aihetta aggregoituna YKR-ruudukkoon. Hakasuluissa solujen lukumäärä.

Englanninkielisiä aiheita kuvaavassa kuvan 42 kartassa on suomenkielisen aineistosta poikkeavasti muutamia kuvioita helposti havaittavissa, joka tekee siitä hieman mielekkäämmän kartan tarkastella. Tämä kuvioiden selkeämpi näkyminen englanninkielisessä aineistossa ei ole yllättävää, sillä aiheiden ajallisissa tarkasteluissa ylempänä kyseisessä aineistossa oli selkeästi muista aiheista selkeästi erottuvia aiheita (kuvat 35 - 40), eikä vastaavanlaisia piirteitä ollut havaittavissa suomenkielisestä aineistosta. Selkeimmin erottuvat

aiheet ovat aamu, ruokailu ja uusivuosi myös spatiaalisessa tarkastelussa. Klusteroituminen ei ole kovin selkeää, tosin uusivuosi on Suomenlinnan alueella laajin yhtenäinen klusteri. Ruokailulla on yhtenäisiä kuvioita, joita ei suoraan voi sanoa yhtenäisiksi klustereiksi esimerkiksi Kaivopuiston, Lauttasaaren ja Kallion alueilla, joiden ulkopuolella sillä on hajanaisempi spatiaalinen rakenne. Aamu-aiheella on selkeitä yhtenäisiä alueita keskustassa Eiran, Töölön ja Kallion alueilla, mutta myös keskustan ulkopuolella Kumpulassa, Malmilla, Itäkeskuksessa ja Vuosaaressa. Uusivuosi-aihe on spatiaalisesti keskittynyt Suomenlinnan saaren lisäksi Käpylään, Jakomäkeen ja Puotilaan, joiden ulkopuolella se on melko tasaisesti dispersoitunut. Viikonloppuaihe on myös varsin dispersoitunut, mutta pieniä keskittymiä on havaittavissa Helsingin pohjoisosissa kuten Puistolassa. Suomalaisuus on myös varsin hajanainen spatiaaliselta rakenteeltaan, mutta selkein keskittymän kaltainen tihentymä on havaittavissa Paloheinän ja Käpylän välistä Helsingin Keskuspuiston alueelta. Juhlinta-aiheen osalta rakenne on myös varsin dispersoitunut lukuun ottamatta Töölön ja Ruoholahden välistä ranta-aluetta. Talvi-aihe on varsin heikosti näkyvä aihe, joka on myös varsin tasaisen hajanainen, mutta pieni keskittymä sijaitsee Haagassa olevan Alppiruusupuiston kohdalla. Ystävät-aiheella ei vaikuta olevan klustereita missään ja sekin on usean muun englanninkielisen aineiston aiheen tavoin varsin dispersoitunut spatiaaliselta rakenteeltaan. Joulu-aihe on myös varsin hajanainen spatiaaliselta rakenteeltaan, tosin kyseistä aihetta on melko paljon yksittäisten ruutujen muodossa Itäkeskuksessa, Lauttasaaressa ja Herttoniemessä.

Samannimisten aiheiden välillä on numeerisia eroja aihekennojen lukumääriä tarkastellessa, esimerkiksi joulu on suomenkielisessä aineistossa yli kaksinkertaisesti suositumpi kuin englanninkielisessä aineistossa ja ruokailu puolestaan on myös yli kaksinkertaisesti suositumpi englannin- kuin suomenkielisessä aineistossa. Melko suuri ero on myös samankaltaisten aihekokonaisuuksien välillä: suomenkieliset sosiaalinen kanssakäynti ja vierailu hallitsevat suurempaa määrää kennoja kuin englanninkieliset ystävät ja juhlinta. Kesä ja talvi ovat vuodenaikoihin liittyviä aiheita, joita on miltei saman verran. Aamu ja aloittaminen ovat merkitykseltään lähekkäisiä aiheita, joiden välillä ei myöskään ole kovin suurta eroa aihekennojen määrässä. Samannimisten ja samankaltaisten aiheiden spatiaalinen päällekkäisyys on melko vaikeasti havaittavissa karttatarkasteluista, mutta sitä vaikuttaa olevan jonkin verran erityisesti sosiaaliseen kanssakäymiseen

liittyvillä aiheilla. Tarkemmassa valittujen kaupunginosien tarkastelussa aiheiden spatiaaliseen päällekkäisyyteen pääsee hieman havainnollisemmin käsiksi.



Kuva 43. Ruokailu-aiheisten julkaisujen etäisyys lähimpään ravintolaan esitettynä kumulatiivisena osuutena etäisyyden mukaan visualisoituna.

Ruokailu-aiheisten julkaisujen sijaintien ja ravintoloiden sijaintien välisen etäisyydellisen hajoamisen (engl. *distance decay*) tarkastelussa kuvassa 43 on nähtävissä, että englanninkieliset ruokailu-aiheen julkaisut ovat keskimäärin hieman lähempänä ravintoloita kuin suomenkieliset ruokailu-aiheen julkaisut. Suomenkielisten ruokailu-aiheisten julkaisujen suurempi etäisyys voi johtua suuremmasta todennäköisyydestä tehdä kyseinen julkaisu käyttäjän kotoa ravintolan sijaan. Puolet ruokailu-aiheisista julkaisuista sijaitsee noin 75 metrin päässä lähimmästä ravintolasta, jonka jälkeen kohdepisteiden etäisyydellinen hajoaminen kiihtyy ja jo kolme neljäsosaa julkaisusta on noin 200 metrin etäisyydellä lähimmästä ravintolasta. Etäisyydellisen hajoamisen kuvaajan välittämä kuva vaikuttaa olevan looginen ottaen huomioon sen, että ravintoloiden sijainnit ovat OpenStreetMap-palvelun sijainteja ja Instagram-julkaisujen sijainti on kohdepisteisiin perustuva eikä todellisiin sijainteihin perustuva. Etäisyyksiä tarkastellessa täytyy muistaa, että Instagram-julkaisuihin liitetyn kohdepistesijainnin ja julkaisun oikean sijainnin välisen virhemarginaalin on todettu olevan noin 20 metriä (Cvetojevic et al. 2016), sekä OSM-kohdepiste voi sijaita hieman eri paikassa

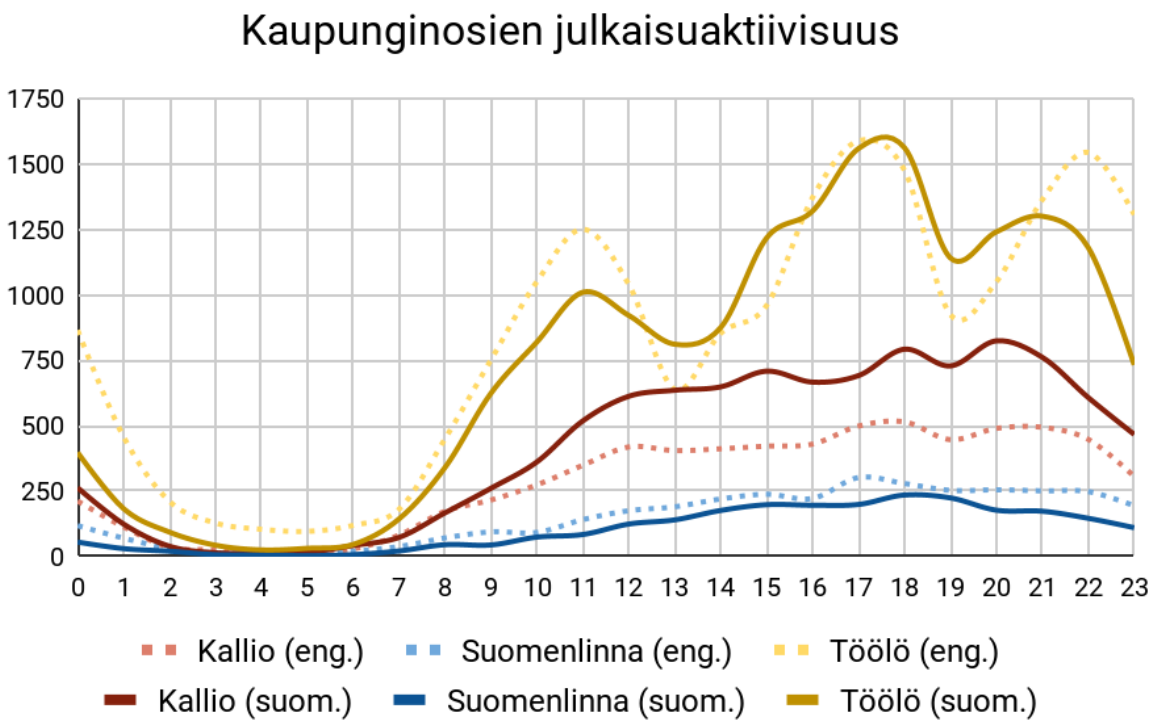
kuin vastaava Instagram-kohdepiste. Syy OSM-kohdepisteiden käyttöön ravintoloiden osalta Instagram-kohdepisteiden sijaan on varsin yksinkertainen. Kuten kappaleessa 2.1 käytiin lävitse, Instagram-aineisto sisältää paljon käyttäjien itse luomia ja nimeämiä kohdepisteitä, joka tuottaa aineistoon useita samannimisiä pisteitä hieman eri sijainneilla. Tämän vuoksi ravintoloiden sijaintitietojen osalta tässä etäisyydellisen hajoamisen tarkastelussa turvauduttiin OpenStreetMap-aineistoon.

3.3.2 Valittujen kaupunginosien tarkastelu

Tarkempaan tarkasteluun valittuja kaupunginosia (Kallio, Töölö ja Suomenlinna) tarkastellessa Instagram-julkaisuista sekä niistä mallinnetuista aiheista paljastuu mielenkiintoisia spatiaalisia ja sisällöllisiä piirteitä. Esimerkiksi julkaisumäärien painopisteiden välillä on selkeitä eroja kieliryhmien välillä ja englanninkielisen aineiston aiheissa näkyy vahvemmin retkeilyyn ja turismiin viittaavia aiheita kuin suomenkielisen aineiston aiheissa. Mielenkiintoinen huomio kaupunginosista mallinnettujen aiheiden koherensseista on se, että muutamien aiheiden osalta pisteytys on yllättävän hyvä verrattuna koko kaupungin aihemallin pisteisiin. Koko kaupungin alueelta mallinnettujen aiheiden koherenssipisteet eivät antaneet syytä odottaa tämän kaltaisia tuloksia. Kenties maantieteellisesti selkeästi kohdennetumpi aineisto pitää myös sisällään kohdennetumpaa ja alueeseen itseensä liittyvää aihesisältöä, jotka täten myös näkyvät vahvemmin alueellisesti toteutetuissa aihemalleissa.

Alueelliset aihemallit toteutettiin lemmatisoiduille kuvateksteille, mutta myös suodatetuille versioille kuvateksteistä, joissa toisessa oli vain substantiiveja ja verbejä ja toisessa oli vain adjektiiveja. Tällä menettelyllä on tarkoitus saada hieman syvyyttä mallinnettaviin aiheisiin mallintamalla aiheita samalta alueelta kolmella eri painopisteellä, mutta myös selvittää, mikäli substantiiveista ja verbeistä koostuvan aineiston mallintaminen kertoisi enemmän Instagram-käyttäjien aktiviteeteistä ja adjektiiveista koostuvan aineiston mallintaminen kertoisi kuvatekstien sävyistä. Ennen näihin tuloksiin syventymistä on todettava, että adjektiiveille tehdyistä aihemalleista on erittäin vaikea tulkita sävyjä ja nimetä syntyneet aiheet niiden sävyä kuvaavin termein adjektiivien samankaltaisuuden vuoksi. Substantiivi-verbi-aihemallin tuloksien erittely oli verrattain varsin helppoa, johtuen osalta maantieteellisesti tiukasti rajatuista alueista, jolloin

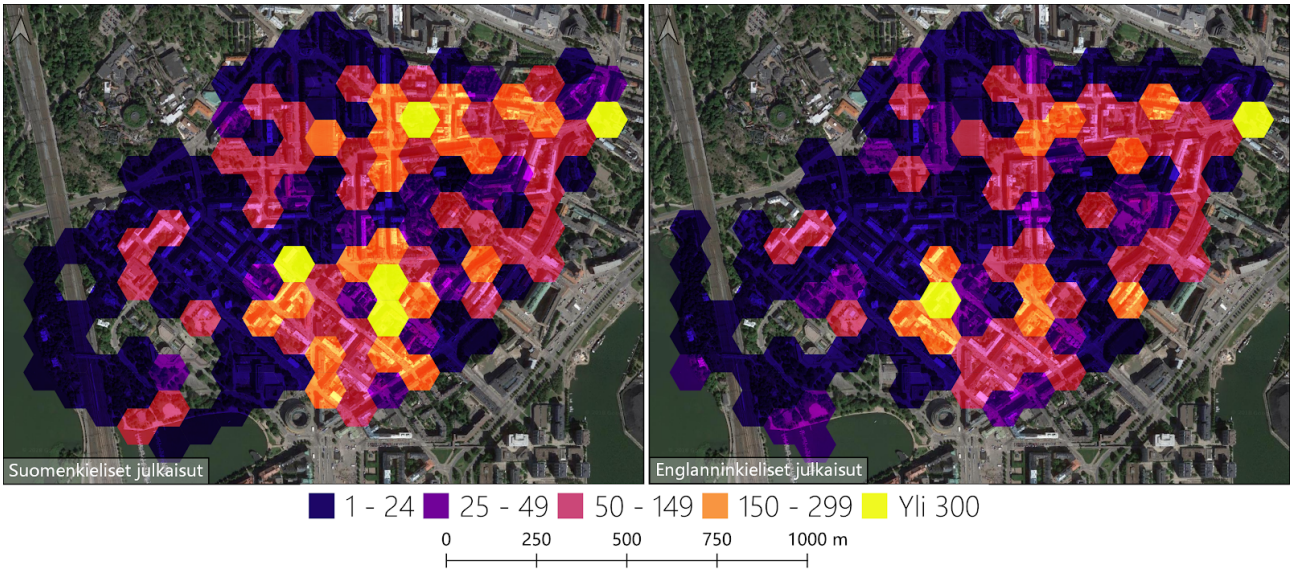
esimerkiksi nähtävyyksiin liittyvät aiheet ovat tunnistettavissa varsin helposti, ja osalta adjektiivien poissulkemisesta joka helpottaa huomattavasti aiheiden tunnistamista ja nimeämistä.



Kuva 44. Kaupunginosien julkaisuaktiivisuus kellonajan mukaan suomen- ja englanninkielisten julkaisujen mukaan eroteltuna. Vaaleammalla sävyllä ja pisteviivalla merkityt ovat englanninkielisiä julkaisuja, tummemmalla yhtenäisellä viivalla merkityt ovat suomenkielisiä.

Kuva 44 antaa kaupunginosista varsin mielenkiintoisen ennakkokuvan, sillä vain Töölön alueella vaikuttaisi olevan aiemmin havaittua kolmipiikkistä rytmiikkaa julkaisuaktiivisuudessa. Kallion alueella on pienemmän piikit vasta iltapäivällä ja illalla, kun taas Suomenlinnassa on vain yksittäinen pehmeä huippu loppuiltapäivästä. Selkeimmät kielten väliset julkaisuaktiivisuus erot ovat Kalliossa, jossa suomenkieliset julkaisut ovat enemmistössä. Töölö ja Suomenlinna ovat julkaisuaktiivisuudeltaan melko samankaltaisia kielten välillä.

3.3.2.1 Kallio



Kuva 45. Instagram-julkaisujen määrät suomen- ja englanninkielisissä aineistoissa Kallion alueelta 100 metrin kennostossa. Taustan satelliittikuva on Googlen satelliittikuva.

Kuvasta 45 on nähtävissä englannin- ja suomenkielisten Instagram-julkaisujen lukumäärät spatiaalisesti esitettynä 100 metrin kuusikulmaiseen kennostoon aggregoituna. Syy 100 metrin kennoston käyttämiseen aluekohtaisissa tarkasteluissa 250 metrin YKR-ruudukon sijaan on varsin yksinkertainen: aiemmin mainittu kuusikulmaisen kennoston parempi soveltuvuus maantieteellisten ilmiöiden tarkasteluun (Birch et al. 2007) sekä 250 metrin YKR-ruudukko vaikutti liian karkealta kaupunginosatason tarkasteluun, jolloin mahdollisesti tärkeä spatiaalinen kuviointi jäisi näkymättömiin. Aluekohtaisissa aihekennovisualisoinneissa (kuvissa 46 - 56) aiheiden aggregointi kennostoon on toteutettu yksinkertaisella dominanssilaskutoimituksella, jossa kennon dominoiva aihe lasketaan sen sisältämän yleisimmän aiheen mukaan. Näin kennokohtaisesti yleisimmät aiheet saadaan visualisoitua kartalle havainnollisella tavalla, mutta yksittäisen kennon luokittuminen jollekin aiheelle ei tarkoita sitä, ettei saman kennon sisällä esiintyisi muita aiheita. Aihekenno siis kuvaa kyseisen alueen yleisintä aihetta. Aihekennojen värityksessä on pyritty seuraamaan tiettyä logiikkaa: sama tai erittäin samankaltainen väritys on valittu samankaltaisille aiheille aineistojen välillä, jotta samankaltaisten aiheiden spatiaalinen vertailu olisi hieman havainnollisempaa.

Suomen kieli dominoi Kallion digitaalista kaupunkitilaa, sillä suomeksi tehtyjä julkaisuja on enemmän ja julkaisut ovat levittäytyneet laajemmalle alueelle, kun taas englanniksi tehdyt julkaisut ovat keskittyneet pienemmälle alalle. Julkaisumäärällisesti intensiivisimpiä paikkoja vertaillessa suomenkielisestä aineistosta

korostuu Helsinginkadun, Vaasankadun ja Sturenkadun muodostama alue, joka ei ole yhtä voimakas englanninkielisessä aineistossa. Lisäksi Kallion kirjaston ja Karhupuiston alue on selkeästi vilkkaampi suomenkielisessä aineistossa. Kallion kirkon alue on molemmissa aineistoissa yhtäläisesti vilkas, kuten on myös Sörnäisten kurvin alue. Nämä löydökset yhdistettynä aiemmin tässä työssä visualisoituun kielidominanssiin (kuva 30) voivat viitata Kallion olevan enemmän paikallisväestön suosima alue, josta tehdään Instagram-julkaisuja suomeksi, sen sijaan, että se olisi esimerkiksi turistien suosima Instagram-julkaisujen alue. Tämä on sinänsä yllättävää, sillä yleinen mielikuva Kalliosta on trendikäs ja kansainvälinen, mutta mielikuva ei vaikuta näkyvän digitaalisessa kaupunkitilassa. Toisaalta täytyy muistaa, että automaattisen kielentunnistuksen jälkeisessä muuten käsittelemättömässä aineistossa englanti on suomea suurempi kieliryhmä. Kielentunnistuksen jälkeen työssä tehdyistä suodatuksista (taulukko 6) johtuen tämä kuvan 45 kartan välittämä tilanne voi olla hieman vääristynyt.

Taulukko 12. Suomenkielisten Instagram-julkaisujen aiheet Kallion alueelta. Aiheiden alapuolella on tärkeysjärjestyksessä aihekohtaiset tärkeimmät sanat. Sanojen jälkeen on aiheen koherenssipisteet, aiheeseen kuuluvien Instagram-julkaisujen lukumäärä ja aihetta kuvaava termi. Malli käsittää kaikki lemmatisoidut sanat.

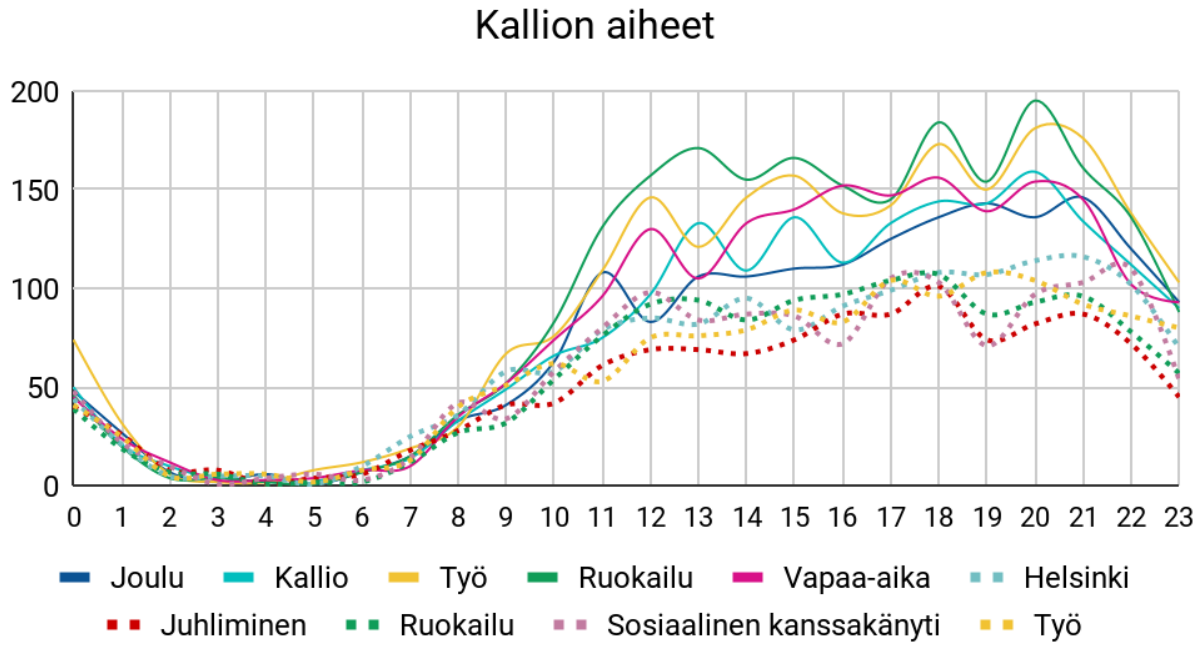
Aihe 1	**Aihe 2**	**Aihe 3**	**Aihe 4**	**Aihe 5**
joulu	kallio	mennä	ihana	päivä
voida	viikonloppu	päivä	paras	tehdä
lähteä	helsinki	aamu	brunssi	kuva
koti	kiva	ilta	seura	kesä
kevät	eka	saada	sunnuntai	kahvi
pikkujoulu	käydä	työ	syödä	ottaa
katu	vähä	viikko	päivä	mieli
hetki	tuoda	päästä	lounas	kaveri
ilta	jäädä	aurinko	puisto	kalja
syksy	kirkko	elämä	ystävä	sanoa
0.25588	***0.12377***	0.25272	**0.47569**	0.19986
1774	1834	2208	2265	1962
Joulu	**Kallio**	**Työ**	**Ruokailu**	**Vapaa-aika**

Taulukko 13. Englanninkielisten Instagram-julkaisujen aiheet Kallion alueelta. Aiheiden alapuolella on tärkeysjärjestyksessä aihekohtaiset tärkeimmät sanat. Sanojen jälkeen on aiheen koherenssipisteet, aiheeseen kuuluvien Instagram-julkaisujen lukumäärä ja aihetta kuvaava termi. Malli käsittää kaikki lemmatisoidut sanat.

Aihe 1	**Aihe 2**	**Aihe 3**	**Aihe 4**	**Aihe 5**
good	party	sunday	friend	day
time	light	love	lunch	work
helsinki	today	day	happy	feel
morning	breakfast	brunch	people	night
life	fun	coffee	yesterday	year
food	snow	saturday	christmas	photo
day	find	kallio	well	start
place	thing	evening	today	will
dinner	shop	great	time	monday
winter	coffee	weekend	finland	office
0.32024	0.27546	0.27140	***0.26416***	0.27969
1495	1197	1360	1392	1384
Helsinki	**Juhliminen**	**Ruokailu**	**Sosiaalinen kanssakäynti**	**Työ**

Taulukkoihin 12 ja 13 on koostettu Kallion alueelta mallinnetut viisi aihetta molemmista kieliaineistoista, aiheiden kymmenen tärkeintä sanaa, koherenssipisteet, aiheeseen kuuluvien julkaisujen lukumäärä ja alimpana kirjoittajan näkemys siitä, mikä aiheen nimi voisi olla. Kuten aikaisemmissa aihemallitaulukoissa, myös tässäkin aiheiden nimet ovat kirjoittajan subjektiivinen näkemys, joten tämän työn lukijoilla voi olla eriäviä mielipiteitä siitä, mikä nimi millekin aiheelle kuuluu. Nimettyjä aiheita tarkastellessa huomaa niiden olevan erittäin samankaltaisia kielten välillä ja itseasiassa kaksi aihetta kummastakin mallista on nimetty identtisesti, “Työ” ja “Ruokailu”. Kallion digitaalisessa kaupunkitilassa keskustellaan molemmilla kielillä työstä ja ruokailusta, jonka perusteella aluetta voisi luonnehtia välttämättömien arjen aktiviteettien kuten työn ja hieman nautinnollisempien aktiviteettien kuten ruokailun osalta melko samanlaisiksi. Ruokailun osalta molemmilla kielillä mallissa tärkeänä sanana on brunssi ja sunnuntai, jotka muodostavat yhdessä todennäköisesti sunnuntaibrunssi-yhdyssanan, jotka ovat varsin suosittuja kyseisessä kaupunginosassa (Stadissa 2018, Karjalainen 2018). Työ-aiheen osalta nimeäminen on onnistunut hieman paremmin englanninkielisten aiheiden osalta, sillä työhön liittyvät sanat ovat hieman tärkeämmillä paikoilla kuin suomenkielisessä aiheessa. Toisaalta suomenkielinen aihe sisältää paljon tärkeitä sanoja, joiden perusteella on varsin vaikea muodostaa hyvää käsitystä kyseisen aiheen nimestä.

Suomenkielisissä aiheissa on kaksi hieman tarkemmin määriteltyä aihetta, “Kallio” ja “Joulu”, kun loput aiheet on nimetty melko suurpiirteisesti. Kallio alueena lienee olevan enemmän suomalaisten Instagram-käyttäjien tai suomeksi kirjoitettujen julkaisujen aiheena, joka kertonee siitä, että se on suomalaisten keskuudessa maineessa oleva alue, kun taas kansainvälisemmällä tasolla Kallio ei näy yhtä erityisenä paikkana digitaalisessa kaupunkitilassa. Tätä tukee myös se, että sana ‘kallio’ esiintyy taulukon 13 kolmannessa aiheessa kuudenneksi tärkeimpänä sanana kyseiseen aiheeseen. Englanninkielisissä aiheissa on yhtenä aiheena Helsinki, joten voi olla, että englanniksi julkaisevat Instagram-käyttäjät eivät erottele kaupunginosia toisistaan vaan he käsittelevät Helsinkiä yhtenä kokonaisuutena. Tämän taustalla voi muun muassa olla matkailijat, jotka eivät välttämättä tunne Helsingin kaupunginosia tai englanniksi julkaisevat suomalaiset, jotka haluavat nostaa Helsingin näkyvyyttä kaupunkina ylipäätään. Juhliminen on vain englanninkieliseen aineistoon mallintunut aihe, jonka perusteella Kallio voi näyttäytyä enemmän juhlimisen ja vapaa-ajanvieton paikkana englanninkielisessä digitaalisessa kaupunkitilassa. Itse asiassa kumpikaan sanoista “juhla” tai “juhliminen” eivät sisälly mihinkään suomenkielisiin aiheisiin taulukossa 10. Joulu mainitaan myös englanninkielisessä aiheessa 4, mutta muiden ja tärkeämpien sanojen osalta kyseinen aihe käsittelee vahvemmin sosiaalista kanssakäymistä. Kielten väliset aihe-erot ovat tässä aihemallissa melko pienet, sillä kielten välillä on kaksi samannimistä aihetta ja kaksi samankaltaista aihetta.



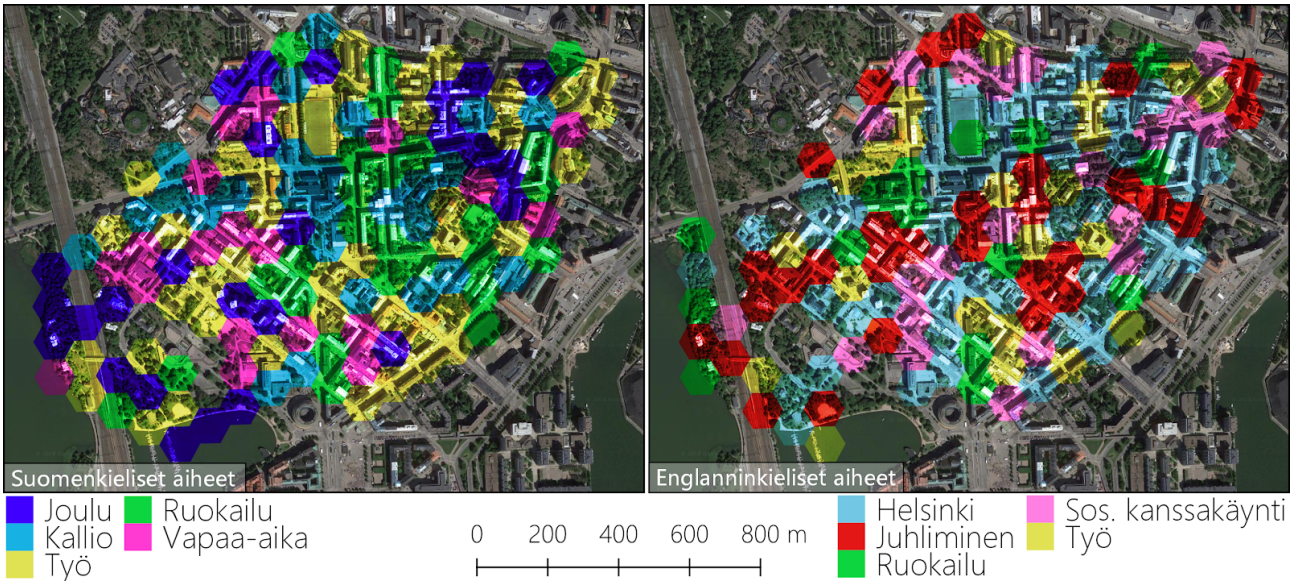
Kuva 46 Kallion suomen- ja englanninkielisten aiheiden ajallinen jakauma kellonajan mukaan. Englanninkieliset julkaisut ovat merkitty pisteviivoin ja samankaltaiset aiheet on väritetty samankaltaisilla väreillä.

Kuvasta 46 paljastuu Kallion alueen kellonaikaan perustuva aiherakenne suomen- ja englanninkielisten julkaisujen välillä. Kolmipiikkinen rytmiikka ei Kallion julkaisuissa juurikaan näy kuin tulkinnanvaraisesti suomenkielisissä aiheissa ja selkeimmät julkaisuhuiput ovat iltaisin klo 18 - 21 välillä. Ruokailu on molemmilla erittäin yleinen aihe läpi koko päivän. Kielten väliset julkaisumääräerot näkyvät kuvaajasta myös hyvin.

Taulukko 14. Ruokailu-aiheisten julkaisujen keskimääräinen etäisyys lähimpään ja viiteen lähimpään ravintolaan Kallion alueella.

	Suom. lähin	**Eng. lähin**	**Suom. 5 lähintä**	**Eng. 5 lähintä**
Keskiarvo	72	68	135	133
Mediaani	56	55	124	117
Keskihajonta	56	55	75	83
1. kvartiili	31	30	78	73
3. kvartiili	109	101	179	178

Ruokailu-aihe esiintyy molemmilla kielillä ja on luonteeltaan sellainen, jonka onnistumisen validointia voi jossain määrin tehdä esimerkiksi katsomalla aiheen julkaisujen keskimääräisiä etäisyyksiä lähimpiin ravintoloihin (taulukko 14). Tällaisessa tarkastelussa selviää englanninkielisten julkaisujen olevan keskimäärin lähempänä ravintoloiden sijainteja myös kaupunginosatasolla, ainakin Kalliossa, koko kaupungin mittakaavan lisäksi. Suomenkielisten julkaisujen keskimääräistä pidempi etäisyys lähimmistä ravintoloista voi viestiä siitä, että suomenkieliset ruokailuun liittyvät julkaisut tehdään hieman useammin esimerkiksi asunnoista, kuten Instagram-käyttäjien kotoa. Ravintoloiden sijainnit ovat kohdepisteitä ja peräisin OpenStreetMapista, jonka vuoksi etäisyysvertailussa on pieni virhetekijä mukana. Esimerkiksi Kalliossa sijaitseva ravintola Du Diin kohdepiste voi sijaita eri koordinaateissa kuin vastaava Instagramiin tallennettu kohdepiste. Tällaisissa tapauksissa etäisyydellinen epävarmuus on melko todennäköisesti vain muutaman metrin luokkaa, mutta epävarmuutta taulukon 14 etäisyyksissä kuitenkin on. Kuten ylempänä kuvan 43 osalta todettiin, ravintoloiden sijainnit ovat OpenStreetMap-pohjaisia, koska puhtaasti ravintoloita sisältävä otanta Instagram-kohdepisteistä on vaikea toteuttaa kattavasti, jolloin epävarmuustekijästä huolimatta OpenStreetMap-aineiston ravintolakohdepisteet ovat yleispiirteisemmällä tasolla luotettavampia.



Kuva 47. Kallion kaikki lemmatisoidut sanat huomioivan aihemallin tulokset aggregoituna 100 metrin kennostoon. Taustan satelliittikuva on Googlen satelliittikuva.

Aihekennoihin aggregoidut englannin- ja suomenkieliset aiheet muodostavat toisistaan eroavan spatiaalisen rakenteen, joka on nähtävissä kuvan 47 aihekennostoissa. Kielten välisiä samankaltaisuuksiakin tosin on, kuten työ- ja ruokailu-aihekennojen sekä Kallio- ja Helsinki-aihekennojen osittainen päällekkäisyys. Työaihekennojen kielten välistä päällekkäisyyttä on linjojen ja Helsinginkadun varrella. Ruokailuaihekennoilla lähinnä Linjojen alueella myös. Kallio- ja Helsinkiaihekennojen päällekkäisyys on laajimmalle levinnyttä.

Suomenkielisissä aiheista laajimmalle levinneet aiheet ovat työ ja Kallio. Työn yleisyys kytkee Kallion suomenkieliset julkaisut arkeen melko selkeästi korostaen Hämeentien vartta, Linjojen aluetta ja Vaasankadun vartta työ-julkaisujen alueina. Suomenkielisen aineiston työ-aihe muodostaa lukuisia yhtenäisiä kahden tai useamman kennon alueita, jotka ovat melko tasaisesti Hämeentien, Linjojen ja Aleksis Kiven kadun varsilla. Torkkelinmäen alueella ei ole suomenkielisessä aineistossa lainkaan työ-aiheen alueita. Ruokailu-aihe muodostaa suomenkielisessä aineistossa painopisteensä itäiselle puolelle Kalliota Fleminginkatua seurailevana pohjois-eteläsuuntaisena vyöhykkeenä, mutta myös Linjojen ja Hämeentien alueilla on muutaman kennon keskittymiä. Näillä alueilla on lukuisia ravintoloita sekä kahviloita. Suomenkielisellä ruokailu-aiheella on vain yksittäisiä päällekkäisiä kennoja englanninkielisen ruokailu-aiheen kanssa lähinnä Linjojen alueella ja Hämeentien varrella. Joulu-aihe on melko laajalle levittäytynyt aihe, joka muodostaa nauhamaisia spatiaalisia kuvioita Tokoinrannassa ja Sörnäisten kurvin lähellä. Kallio-aihe

muodostaa kolme isohkoa keskittymää keskelle Kalliota Helsinginkadun varteen ja eteläpuolelle. Tämän lisäksi Kallio-aihetta on hajanaisesti ympäri aluetta esimerkiksi Karhupuiston, Kallion virastotalojen ja Vaasanpuistikon eli niin sanotun “Piritorin” ympärillä. Nämä alueet voivat olla Instagram-käyttäjien mielestä Kalliota hyvin edustavia alueita. Suomenkielisistä viimeinen aihe, vapaa-aika, vaikuttaa olevan vahvasti länsipainotteinen, sillä suurin osa vapaa-aika-aiheen kennoista sijaitsee linjojen alueella. Itseasiassa, jokaisen vapaa-aika-aiheen kennon vieressä on työ-aiheen kenno, jonka myötä ainakin Instagram-julkaisujen kuvatekstien kautta näyttäytyvässä digitaalisessa kaupunkitilassa Kalliossa työ- ja vapaa-ajan alueet ovat lähekkäin.

Englanninkielistä aihekennostoa hallitsee ensimmäinen aihe, Helsinki, lähes kauttaaltaan. Kenties Kallio edustaa Helsinkiä melko hyvin englannin kielellä julkaisevien Instagram-käyttäjien mielestä. Selkein Helsinki-aiheen keskittymä on Helsinginkadun ja Aleksis Kiven kadun väliin jäävällä alueella. Helsinki- ja Kallio-aiheet ovat samankaltaisia paikkaan liittyviä aiheita, jonka vuoksi niillä on sama väritys kartassa. Seuraavaksi yleisin aihe, juhliminen, vaikuttaa muodostavan itä-länsisuuntaisen käytävän, joka halkoo Kalliota Ensi linjan ja Eläintarhantien risteyksestä Torkkelinmäellä saakka. Kalliossa on melko paljon ravintoloita- ja baareja näillä alueilla, joten tämä ei varsinaisesti ole yllättävää. Juhlimisella on myös muutamia yksittäisiä kennoja Kallion reuna-alueilla, kuten Sörnäisten kurvissa ja Tokoinrannassa. Työ-aiheen osalta kennoja on melko vähän ja päällekkäisyyttä suomenkielisen työ-aiheen kanssa on vain hieman, esimerkiksi Sörnäisten kurvin, Hämeentien ja Helsinginkadun varrella, mutta ne ovat yksittäisiä kennoja. Englanninkielisen työ-aiheen suurin keskittymä on Brahenkentän länsipuolella, kun taas suomenkielisellä työ-aiheella on useampi iso keskittymä. Ruokailun osalta englanninkielisessä aineistossa on lähinnä yksittäisiä kennoja, mutta Siltasaarenkadun, Töölönlahden itärannan ja Tauno Palon puiston alueet ovat selkeitä ruokailu-aiheen alueita, joilla sijaitsee myös ravintoloita ja kahviloita. Sosiaalinen kanssakäynti-aihe muodostaa nauhamaisia alueita Linjojen, Sörnäisten kurvin ja Porvoonkadun varsille, mutta on muuten melko dispersoitunut.

Taulukko 15. Suomenkielisten julkaisujen substantiivien ja verbien perusteella muodostettu aihemalli Kallion alueelta.

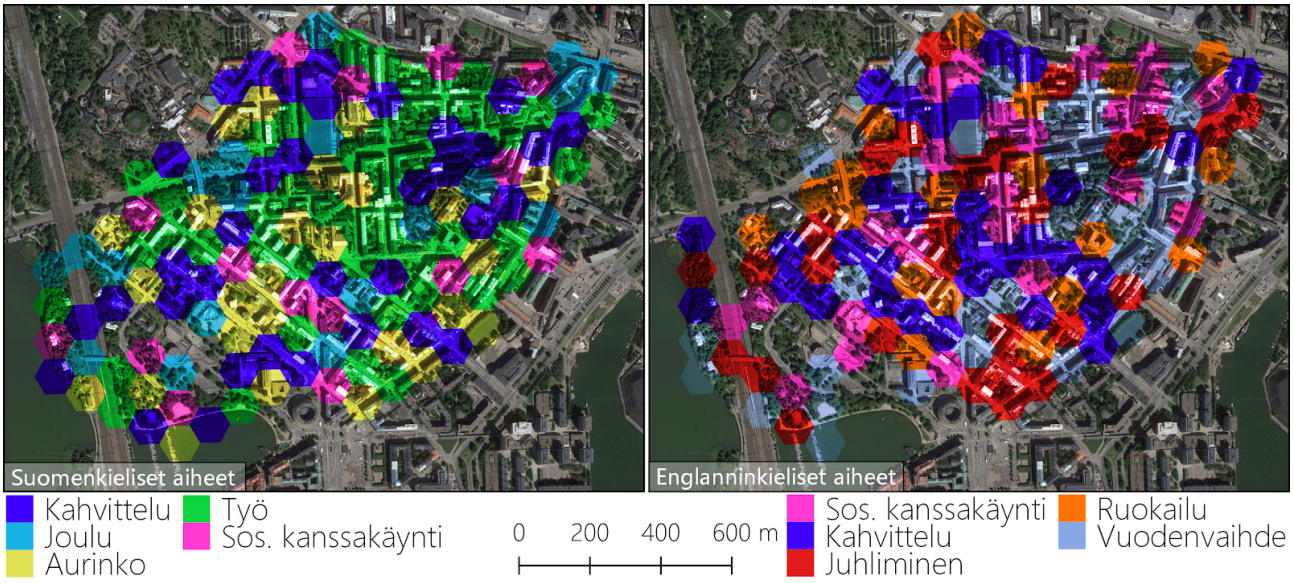
Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
ilta	koti	päivä	kallio	mennä
kesä	joulu	aurinko	tehdä	saada
kahvi	voida	aamu	pitää	ihminen
helsinki	kevät	huomen	työ	seura
elämä	brunssi	näyttää	lähteä	hetki
sunnuntai	vetää	saada	käydä	alkaa
syödä	saada	fiilis	kuva	ruoka
lauantai	keikka	tunnelma	voida	nainen
karhupuisto	naapuri	paistaa	päästä	tuoda
kaupunki	vapaapäivä	tehdä	viikonloppu	käydä
0.21187	0.3243	0.26069	0.31761	**0.33188**
1909	1663	1782	2156	1805
Kahvittelu	**Joulu**	**Aurinko**	**Työ**	**Sosiaalinen kanssakäynti**

Taulukko 16. Englanninkielisten julkaisujen substantiivien ja verbien avulla muodostettu aihemalli Kallion alueelta.

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
yesterday	today	helsinki	time	day
friend	coffee	finland	lunch	love
friday	day	find	night	morning
cake	food	party	kallio	year
place	life	brunch	today	will
winter	dinner	summer	thing	sunday
sun	photo	night	guy	work
town	work	kallio	eat	feel
light	saturday	beer	buy	christmas
window	evening	view	drink	time
0.17634	0.27735	0.2933	***0.14899***	**0.3399**
1150	1373	1342	1215	1444
Sosiaalinen kanssakäynti	**Kahvittelu**	**Juhliminen**	**Ruokailu**	**Vuodenvaihde**

Pelkästään Instagram-julkaisujen kuvatekstien lemmatisoituja substantiiveja ja verbejä hyödyntäneet aihemallit on koostettu taulukoihin 15 ja 16. Verrattuna kaikkia lemmatisoituja sanoja hyödyntäneitä malleja (taulukot 12 ja 13) näihin malleihin voi havaita melko paljon samankaltaisuutta, joka tosin pieneni hieman, ja enemmän toistaan selkeästi eroavia aiheita. Adjektiivien jääminen kokonaan pois tästä mallinnuksesta tekee tuloksista tarkemman oloiset ja helpottaa erityisesti aiheiden nimeämistä huomattavasti. Suomenkieliset aiheet muuttuivat hieman, mutta suurin osa aiheista pysyi periaatteessa samoina vaikkakaan

eivät säilyttäneet paikkaansa järjestyksessä. Aiheista sosiaalinen kanssakäynti onnistui parhaiten ja kahvittelu huonoiten. Täysin uusi aihe, aurinko, joka ei näkynyt taulukon 12 aiheissa, mallintui pelkästään substantiiveja ja verbejä hyödyntävään aineistoon. Englannin osalta tapahtui enemmän muutoksia, sillä kaksi uutta aihetta, kahvittelu ja vuodenvaihde, mallintuivat. Näistä vuodenvaihde on koherenssipistein mitattuna onnistunein aihe. Kyseisen aiheen nimeämisen vuodenvaihde-nimiseksi voi kaivata hieman perusteluja. Aiheelle tärkeitä sanoja ovat “year” ja “christmas”, jotka tuovat nopeasti mieleen joulun ja uuden vuoden aika erityisesti, kun katsoo muita tärkeitä sanoja. Lisäksi tulee muistaa että “new”-sana on poistettu hukkasanana aineistosta. Englanninkielinen kahvittelu-aihe voisi myös olla nimellä ruokailu, mutta koska sana “coffee” on aiheelle tärkeämpi sana kuin “food” ja “dinner” aihe liittynee lähemmin kahvista nauttimiseen kuin ruokailuun. Sama pätee myös suomenkieliseen kahvittelu-aiheeseen. Englanninkielisen aineistoin huonoimmin onnistunut aihe on ruokailu koherenssipistein mitattuna. Kuten kaikki sanat huomioonottavassa mallissa, myös substantiivi-verbimallissa on aihe-erojen osalta samankaltaisia suhteita: kielten välillä on kaksi samannimistä aihetta, mutta vain yksi samankaltainen aihe, toisistaan selkeästi eroavia aiheita on kaksi.



Kuva 48. Kallion alueen substantiiveihin ja verbeihin perustuvan aihemallin tulokset aggregoituna 100 metrin kennostoon. Taustan satelliittikuva on Googlen satelliittikuva.

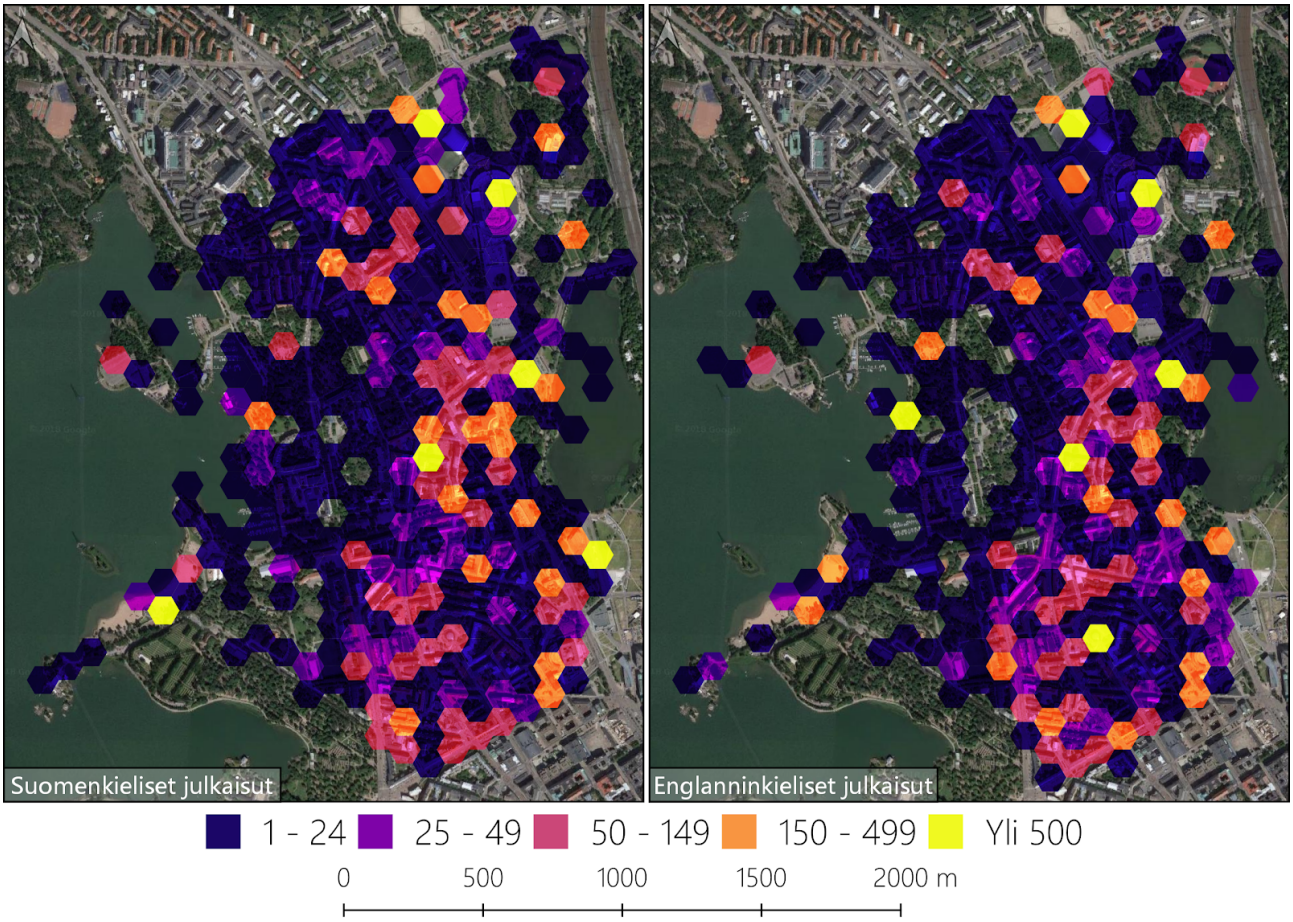
Kuvan 48 karttaan on visualisoitu aiheet, jotka on mallinnettu pelkästään kuvatekstien substantiivien ja verbien perusteella (taulukot 14 ja 15). Päällekkäisiä aihekennoja ovat kahvittelu, sosiaalinen kanssakäynti sekä joulu/vuodenvaihde. Kahvittelun osalta selkeimmät päällekkäisyydet ovat linjojen ja Brahenkentän

ympäristössä. Sosiaalinen kanssakäynti on päällekkäistä lähinnä Aleksis Kiven kadun varrella. Joulun ja vuodenvaihteen aihekennojen päällekkäisyys toteutuu Linjojen, Torkkelinmäen ja Brahenkentän alueella.

Suomenkielisiä aihekennoja tarkastellessa selviää, että työ-aihe on selkeästi laajimmalle levittäytynyt aihe. Se muodostaa laajahkon yhtenäisen ja sormimaisen alueen Neljännen linjan ja Aleksis Kiven kadun väliselle laajalle alueelle. Verrattuna kuvan 47 suomenkieliseen työ-aiheeseen tässä alueellisessa aihemallissa se on laajemmalle levittäytynyt, vaikka aiheeseen luokiteltuja julkaisuja on hieman vähemmän. Työ-aiheen jälkeen kahvittelu-aihe on seuraavaksi hallitsevin aihe, jolla on neljä suurempaa keskittymää: Brahen kentän pohjoispuoli, Helsinginkadun varrella sekä kaksi keskittymää Toisen linjan varrella. Aurinko- ja joulu-aiheet ovat yhtä yleisiä spatiaalisesti, tosin aurinko on eteläpuolelle painottunut, kun joulu on taas länsipuolelle painottunut. Joulu-aihe säilyi aiheena myös substantiiveja ja verbejä hyödyntävässä aihemallissa ja kuvia 49 sekä 50 vertailemalla, aiheen spatiaalinen rakenne on osittain päällekkäinen. Sosiaalinen kanssakäynti on suomenkielisistä aiheista spatiaaliselta rakenteeltaan hajautunein.

Englanninkielisiä aihelevinneisyyksiä tarkastelemalla (kuva 48) näkee, että vuodenvaihde-aihe on laajimmalle levittäytynyt aihe. Se muodostaa suurehkon yhtenäisen keskittymän Kallion koillisnurkkaan. Seuraavaksi laajimmalle levittäytynein aihe on kahvittelu, joka on myös tunnistettu suomenkielisestä aineistosta, mutta on laajemmalle levittäytynyt englanninkielisessä. Sillä on viivamaisia alueita linjojen ja Karhupuiston alueella. Loput kolme englanninkielistä aihetta (sosiaalinen kanssakäynti, juhliminen ja ruokailu) ovat spatiaalisesti yhtä laajalle levittäytyneitä. Sosiaalinen kanssakäynti on selkeästi vahvempi kuin suomenkielisen aineiston vastaava ja muodostaa useamman keskittymän. Juhliminen muodostaa usean pienehkön keskittymän Helsinginkadun varrelle ja Sörnäisten kurviin. Lisäksi sillä on keskittymiä Siltasaarenkadun eteläpäädyssä ja Tokoinrannassa. Englanninkielinen vuodenvaihde-aihe on melko lähellä suomenkielistä joulu-aihetta ja niillä onkin muutamia spatiaalisia päällekkäisyyksiä, mutta vuodenvaihde on huomattavasti laajemmalle levittäytynyt kuin suomenkielinen joulu-aihe.

3.3.2.2 Töölö



Kuva 49. Instagram-julkaisujen määrät suomen- ja englanninkielisissä aineistoissa Töölön alueelta 100 metrin kennostossa. Taustan satelliittikuva on Googlen satelliittikuva.

Töölössä Temppeliaukion kirkko, Regatta-kahvila ja Sibelius-monumentti korostuvat englanninkielisessä aineistossa verrattuna suomenkieliseen, kun tarkastellaan julkaisumäärien spatiaalista levinneisyyttä, joka kertonee aineistossa olevan jonkin verran turisteja. Temppeliaukion kirkko ja Sibelius-monumentti ovat eräitä Helsingin tärkeimpiä nähtävyyksiä. Molemmilla kielillä korostuvat Oopperatalo, Olympiastadion ja Helsingin Jäähalli. Suomenkielisessä aineistossa englanninkielisestä poiketen korostuvat lisäksi Hietalahden uimaranta ja Finlandia-talo, Hietalahden uimaranta on jonkin verran korostunut myös englanninkielisessä aineistossa. Molemmissa aineistoissa vaikuttaa olevan Runeberginkadun vartta seuraileva julkaisujen alue, joissa tihentymät ovat etelässä Kampin lähellä ja pohjoisempana Mannerheimintien ja Runeberginkadun risteyksessä. Molempia aineistojen spatiaalista levinneisyyttä yhdistävät myös vähäiset julkaisumäärät länsi- ja pohjoisosissa Töölöä, jotka ovat pääosin asuinaluetta ja läheisten sairaaloiden kiinteistöjä.

Taulukko 17. Suomenkielisten Instagram-julkaisujen aiheet Töölön alueelta. Aiheiden alapuolella on tärkeysjärjestyksessä aiheiden kannalta tärkeimmät sanat. Sanojen jälkeen on aihekohtaiset koherenssipisteet, aiheeseen kuuluvien Instagram-julkaisujen lukumäärä ja aihetta kuvaava termi

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
kuva	ilta	mennä	ihana	päivä
pitää	päivä	helsinki	paras	kesä
saada	peli	lähteä	päivä	tehdä
nähdä	viikko	käydä	joulu	voida
iso	alkaa	koti	seura	suomi
tehdä	ooppera	paikka	aamu	eka
the	helsinki	kohta	viikonloppu	käydä
voida	kausi	sunnuntai	kiva	kuva
maali	keikka	kaunis	työ	kevät
katsoa	lenkki	loma	ystävä	vähä
0.3933	***0.2520***	0.28675	0.31425	0.30851
3636	3728	3341	3959	3628
Valokuvaus	**Tapahtuma**	**Helsinki**	**Joulu**	**Kesäpäivä**

Taulukko 18. Englanninkielisten Instagram-julkaisujen aiheet Töölön alueelta. Aiheiden alapuolella on tärkeysjärjestyksessä aiheiden kannalta tärkeimmät sanat. Sanojen jälkeen on aihekohtaiset koherenssipisteet, aiheeseen kuuluvien Instagram-julkaisujen lukumäärä ja aihetta kuvaava termi.

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
year	game	time	love	good
happy	team	morning	helsinki	helsinki
day	fun	day	will	work
finland	canada	walk	church	day
finnish	guy	evening	rock	beautiful
light	friend	finally	christmas	today
autumn	hockey	sunday	photo	great
sibelius	play	perfect	day	finland
sweden	party	lovely	time	ready
girl	win	place	open	summer
0.31847	**0.49077**	0.29785	0.3408	***0.23195***
4102	3401	4502	3631	5147
Suomi	**Urheilu**	**Kävely**	**Temppeliaukion kirkko**	**Helsinki**

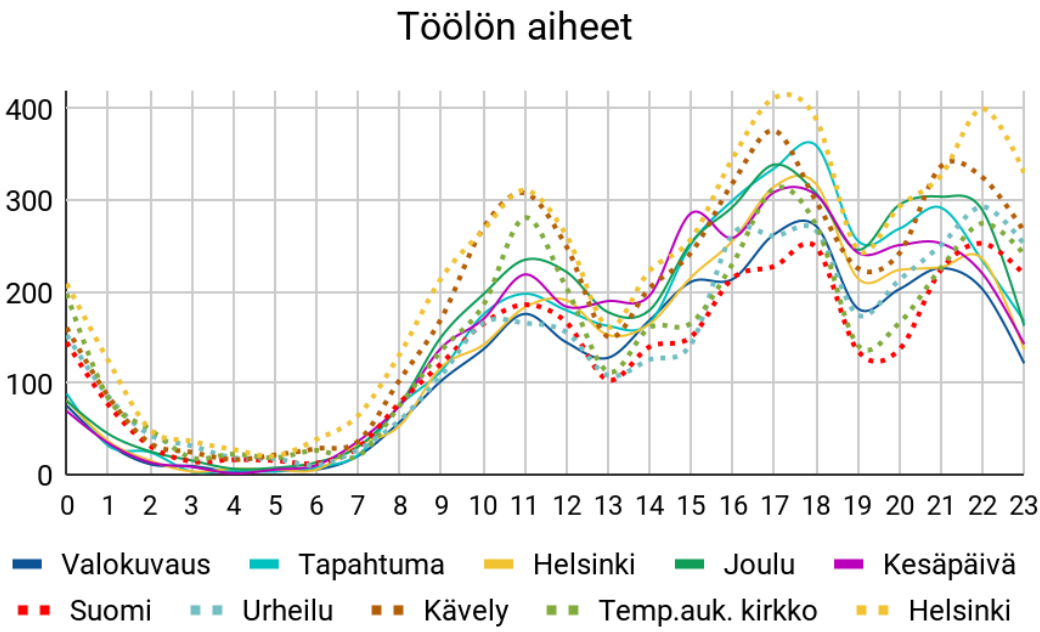
Kun Töölön alueelta mallinnettuja aiheita tarkastelee taulukoista 17 ja 18, voi huomata aiheiden olevan melko lähellä toisiaan, sekä hieman tarkemmin määriteltyjä kuin koko Helsingin mittakaavassa tehdyssä aihemallinnuksessa kuten huomattiin myös Kallion alueaihemallien kanssa. Molemmissa aihemalleissa on mallinnettu samannimisiä aiheita yksi, Helsinki, ja samankaltaisia aiheita: tapahtuma suomen- ja urheilu

englanninkielisessä aineistossa. Aihe-erojen osalta Töölö on siis selkeästi Kalliota erilaisempi kielten välillä. Samankaltaisten aiheiden, tapahtuman ja urheilun, sanaston lähemmässä tarkastelussa huomaa, että suomenkielinen tapahtuma-aihe sisältää urheilu- ja kulttuuritapahtumiin liittyviä sanoja. Englanninkielisen urheilu-aiheen sanat liittyvät selkeämmin urheiluun eivätkä sisällä muunlaisiin tapahtumiin kovin vahvasti liittyviä sanoja. Erityisesti huomioitavaa on urheilu-aiheen koherenssipisteytyksen korkeus, joka on kaikkien aihemallien kesken toiseksi korkein koherenssipisteytys. Korkea koherenssipisteytys näkyy myös suoraan sanojen kautta, jotka selkeästi liittyvät urheiluun ja erityisesti jääkiekkoon.

Helsinki on myös molemmissa aineistossa ilmenevä aihe ja tärkeimmiltä sanoiltaan varsin samankaltainen kielten välillä. Selkein ero mallien välillä on varsin erityisen töölöläisen paikan, Temppeliaukion kirkon (englanniksi usein muodossa *rock church*), näkyminen englanninkielisessä mallissa ja sen puuttuminen suomenkielisestä aineistosta. Englanninkielisessä aineistossa on Helsingin ja Temppeliaukion kirkon lisäksi aiheena Suomi ylipäätään. Tämä tukee ajatusta siitä, että suomenkieliset julkaisut liittyvät enemmän arkeen ja arkiseen vapaa-aikaan, kun taas englanninkieliset julkaisut voivat värittyä matkailun sekä kenties hieman suomalaisten kotipaikkaylpeydestä. Kotipaikkaylpeys saatetaan ilmaista englannin kielellä, jotta käyttäjä saa julkaisulleen laajemman yleisön ja tällöin omaa arkista ympäristöä voidaan pyrkiä katsomaan esimerkiksi matkailijan silmin.

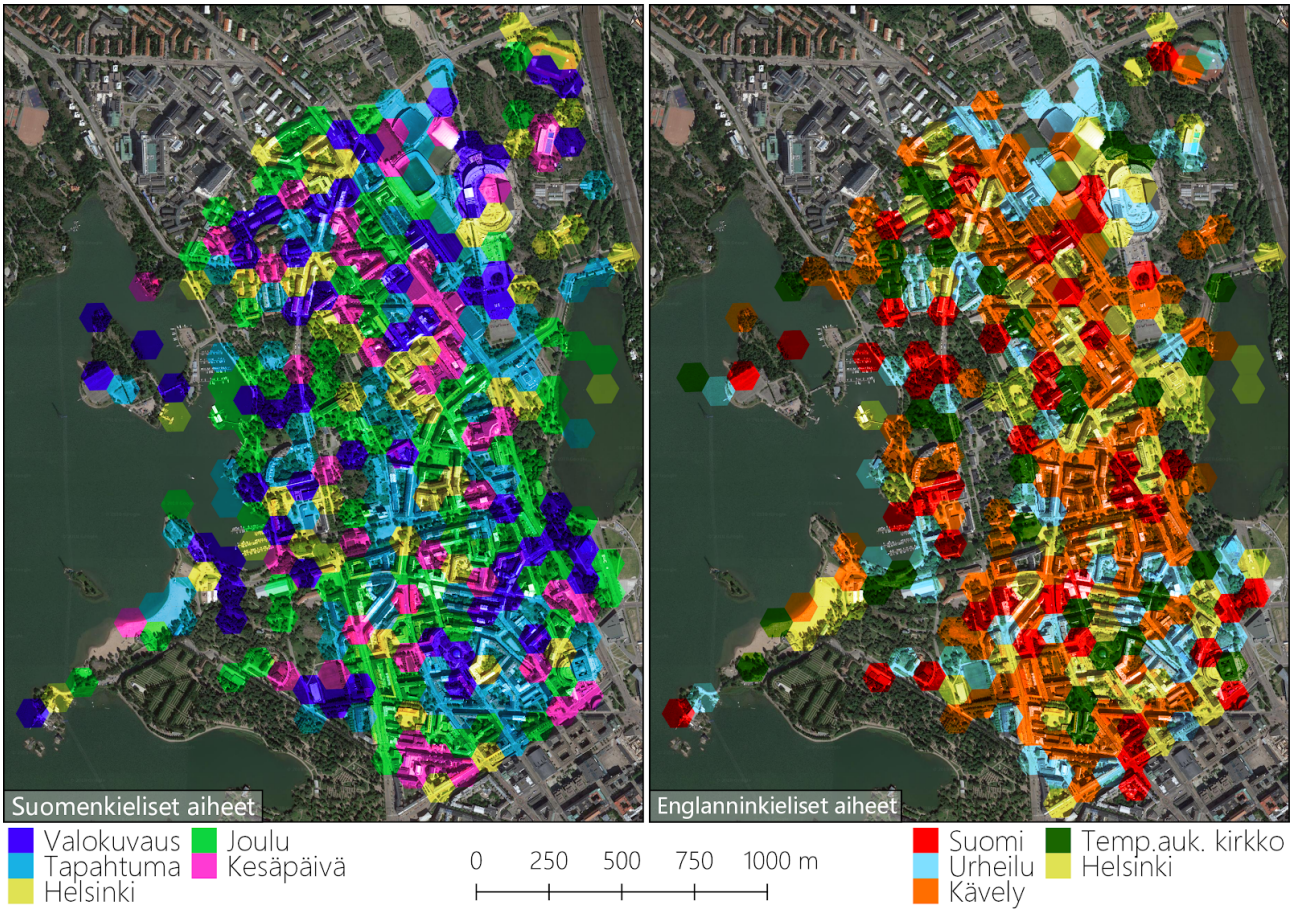
Suomenkielisissä aiheissa näkyy Kallion suomenkielisten aiheiden kanssa yhteneväisesti (taulukko 12) myös joulu. Joulu-aiheen nimeäminen oli ongelmallista, sillä aiheen tärkeimmät sanat ovat adjektiiveja, jotka eivät informoi aiheen osalta juuri mitään ja muut sanat liittyvät vahvasti sosiaaliseen kanssakäymiseen ja viikonaikoihin. Aiheen nimeäminen jouluksi alleviivaa hyvin aiheiden nimeämisen subjektiivisuutta, sillä sanat muodostavat melko ambivalentin kokonaisuuden, jossa kaikki muut sanat eivät ole suoranaisesti jouluun liittyviä vaan voivat liittyä paljon epämääräisempiin ja arkisempiin hetkiin sekä aikoihin. Mielenkiintoista on lisäksi se, että aiheisiin on mallintunut joulun kanssa vuodenajallisesti katsottuna täysin vastakkainen aihe, kesäpäivä. Vaikka aiheista valokuvaus on saanut parhaimman koherenssipisteytyksen suomenkielisten aiheiden osalta, aiheen nimeäminen oli haasteellista, sillä tärkeimmän sanan jälkeisten

sanojen perusteella on vaikea muotoilla selkeää aihepiiriä. Tämän vuoksi aihe nimettiin pitkälti tärkeimmän sanan kautta muodostettuna.



Kuva 50. Töölön suomen- ja englanninkieliset aiheet kellonajan mukaan. Englanninkieliset aiheet on merkitty pisteviivein ja samankaltaisilla aiheilla on samanlainen väritys. Kolmipiikkinen rytmiikka näkyy Töölön aineistossa selkeästi.

Kuvasta 50 Näkee Töölön aiheiden temporaalisen rakenteen, joka ei vaikuta eroavan kovin paljoa koko Helsingin temporaalisesta julkaisu- tai aiherakenteesta (kuvat 9, 35, 36). Englanninkieliset Helsinkiä ja kävelyitä käsittelevät aiheet hallitsevat lukumääräisesti Töölössä, suomenkielisistä aiheista taas kesäpäivä ja tapahtuma ovat melko yleisiä. Ajassa tarkasteltuna aiheet vaikuttavat seurailevan samankaltaista rytmiä keskenään eikä juuri mikään aihe erotu joukosta selkeästi erilaisena rytmiltään, tosin Temppeliaukion kirkko saa aamulla ainoan massasta erottuvan piikkinsä ja kesäpäivällä on iltapäivällä kaksi piikkiä. Englanninkieliset aiheet ovat selkeästi yleisempiä myöhäisinä aikoina kuten klo 22-02 välillä, jolloin niitä on peräti kaksinkertaisesti verrattuna suomenkielisten aiheiden julkaisuihin. Tämä aika ero voi viitata useaan otteeseen mainittuun turismin vaikutukseen, sillä turistit ovat matkallaan usein vapaa-ajan vietossa, joten aktiivinen aika päivässä voi venyä myöhemmäksi kuin arkena.



Kuva 51. Töölön alueen kaikki lemmatisoidut sanat huomioonottavan aihemallinnuksen dominoivat aiheet aggregoituna 100 metrin kennostoon. Taustan satelliittikuva on Googlen satelliittikuva.

Kuten on huomattavissa kuvasta 51, Töölön aihemaisemassa molemmilla kielillä on samankaltaista spatiaalista rakennetta päällekkäisten aiheiden osalta melko paljon. Usealla aiheella on selkeitä keskittymiä eri puolilla kaupunginosaa, mutta osa aiheista on selkeästi hajanaisempia. Molemmilla kielillä samankaltainen aihe, tapahtuma ja urheilu, on merkitty karttakuvaan turkoosilla. Urheilu-aiheen selkein keskittymä on Olympiastadionin ja Helsingin jäähallin alueella, mutta aihe hallitsee muitakin alueita kuten esimerkiksi Soutustadionia, Uimastadionia, Eläintarhan urheilukenttää, Taivallahden tenniskeskusta ja Väinämöisen jalkapallokenttää. Tapahtuma-aiheella on päällekkäisyyksiä urheilu-aiheen kanssa niin spatiaalisesti kuin sisällöllisestikin, kuten ylempänä mainittiin. Olympiastadionin ja jäähallin alueella on laaja-alaista päällekkäisyyttä ja jonkin verran myös Soutustadionin alueella. Muuten tapahtuma-aiheessa korostuu esimerkiksi oopperatalon ja kisahallin välinen alue, suuri osa eteläpäätyä Töölöstä ja Runeberginkadun varsi Keski-Töölössä. Molempien aineistojen aihemalleihin oli mallinnettu myös toisiaan täysin vastaava aihe, Helsinki, joka on englanninkielisessä aineistossa laajemmalle levittäytynyt kuin

suomenkielisessä aineistossa. Molemmilla kielillä Helsinki-aihetta on merkitty myös isoimmille urheilupaikoille. Suomenkielistä Helsinki-aihetta on merkitty myös satunnaisemmille paikoille kuten Töölön sairaalan ympäristöön, Hietaniemen niemennokkaan ja Hesperian Esplanadin varrella. Englanninkielinen Helsinki-aihe keskittyy selkeästi kahteen kohtaan: Mannerheimintien ja Helsinginkadun risteyksen ympäristö ja Arkadiankadun varsi. Lisäksi Helsinki-aiheen alaiseksi on merkitty suuri osa Hietaniemen uimarannasta englanninkielisessä aineistossa.

Suomenkielistä aihemaisemaa tarkastellessa ensimmäinen aihe, valokuvaus, on melko laajalle levinnyt ja kyseisen aiheen alle merkityt alueet ovatkin varsin kuvauksellisissa kohteissa kuten rantojen varsilla, nähtävyyksien lähellä ja tapahtumapaikkojen alueella. Valokuvauksen alle on merkitty esimerkiksi Sibelius-monumentin paikka, Temppeliaukion kirkko ja Kansallismuseon ympäristö sekä useat urheilupaikat kuten Olympiastadion ja Eläintarhan urheilukenttä. Joulu-aihe suomenkielisessä aineistossa on spatiaalisesti varsin laajalle levinnyt, vaikka kuten hieman ylempänä mainittiin, kyseisen aiheen nimeäminen jouluksi oli hankalaa. Kyseisen aiheen spatiaalinen levinneisyys siis johtuu osaltaan siitä, että aiheen alle on luokittunut julkaisuja joiden kuvateksteissä esiintyvät muut tärkeimmät sanat, mutta ei joulu-sanaa. Vuodenajallisesti täysin päinvastainen aihe, kesäpäivä, ei ole yhtä laajalle alueelle levinnyt kuin joulu, mutta muodostaa mielenkiintoista spatiaalista kuviointia Taka-Töölön alueella Mannerheimintien varteen Kisahallin kohdalle. Lisäksi kesäpäivä-aihe on esillä uimastadionilla, Olympiastadionilla, Hietaniemenkadun ja Arkadiankadun varrella.

Englanninkielisessä aineistossa Suomi-aihe vaikuttaa seurailevan osittain samoja alueita kuin suomenkielisen aineiston valokuvaus-aihe. Se ei juuri muodosta suuria keskittymiä, mutta se muodostaa pienempiä keskittymiä Sibelius-monumentin luona, Suomen ympäristökeskuksen rakennuksen vieressä, Toivo Kuulan puiston vieressä ja Humalistonkadun varressa. Kävely-aihe on suurin englanninkielinen aihe spatiaalisesti tarkasteltuna ja muodostaakin suuria yhtenäisiä alueita läpi koko Töölön alueen. Temppeliaukion kirkko-aihe käsittää sisälleen Temppeliaukion kirkon, mutta myös muita alueita Töölössä. Aiheen nimeäminen Temppeliaukion kirkoksi oli ongelmallista, kuten ylempänä mainittiin, sillä muut aiheen

sanat (taulukko 18) käsittelevät muita aiheita, kuten Helsinkiä ja joulua, joten on oletettavissa, että osassa aiheen alle merkityistä julkaisuista aiheena ovat nämä eivätkä niinkään Temppeliaukion kirkko.

Taulukko 19. Suomenkielisten substantiivien ja verbien perusteella muodostettu viiden aiheen aihemalli Töölön alueelta.

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
mennä	saada	helsinki	hetki	päivä
seura	ilta	käydä	tyttö	kesä
ystävä	alkaa	viikonloppu	tehdä	suomi
voida	kuva	viikko	saada	peli
joulu	nähdä	ooppera	lapsi	maailma
aamu	aurinko	päästä	helsinki	lähteä
koti	käydä	keikka	jatkua	voitto
yö	ottaa	ihminen	kahvi	syödä
mieli	syksy	paikka	meno	katsoa
tehdä	voida	työ	klo	matsi
0.25395	0.24213	***0.19797***	**0.30836**	0.2032
3474	3456	3277	2872	3508
Sosiaalinen kanssakäynti	**Valokuvaus**	**Tapahtuma**	**Perhehetki**	**Urheilu**

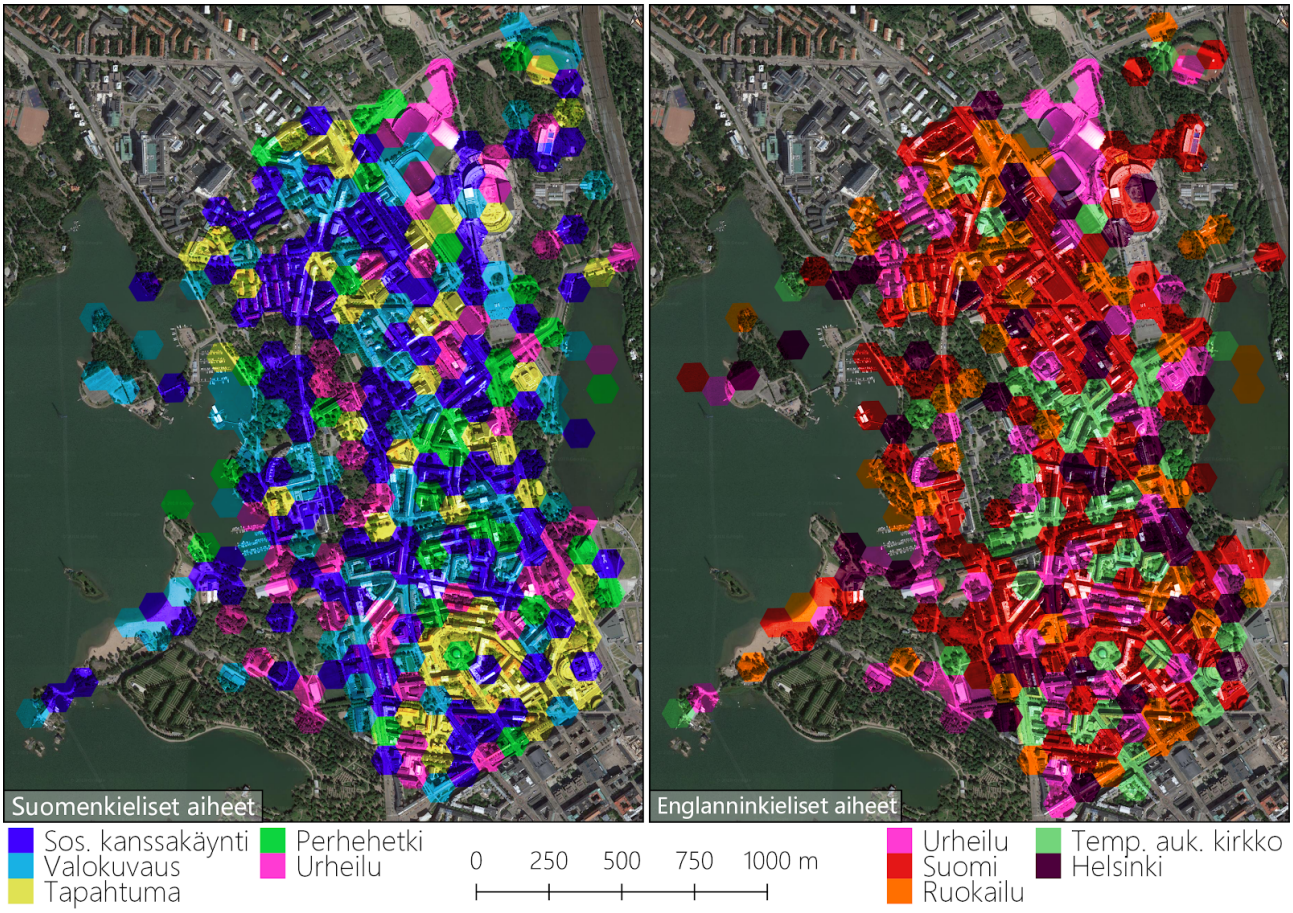
Taulukko 20. Englanninkielisten substantiivien ja verbien perusteella muodostettu viiden aiheen aihemalli Töölön alueelta.

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
work	time	night	day	helsinki
game	finland	guy	church	today
photo	year	sunday	rock	morning
today	love	winter	helsinki	city
team	summer	love	yesterday	christmas
evening	life	watch	friend	view
will	place	light	play	walk
wait	people	dinner	girl	week
win	thing	bit	autumn	weekend
start	will	eat	birthday	sun
0.37699	0.28881	***0.19010***	0.32821	0.22465
3329	4494	3274	3584	4289
Urheilu	**Suomi**	**Ruokailu**	**Temppeliaukion kirkko**	**Helsinki**

Pelkästään substantiiveja ja verbejä hyödyntäneen aihemallinnuksen Töölön alueelle merkityistä Instagram-julkaisuista tulokset on koostettu taulukoihin 19 ja 20. Kielten väliset aihe-erot kasvoivat, sillä samannimisiä aiheita on vain yksi, urheilu, eikä samankaltaisia aiheita juurikaan ole. Aiheet muuttuivat

suomenkielisen aineiston osalta jonkin verran, kun taas englanninkielisen aineiston osalta aiheet eivät ole muuttuneet kovin paljoa kaikki lemmatisoidut sanat huomioineista aihemalleista. Suomenkielisessä aineistossa on tapahtunut enemmän muutoksia esimerkiksi tapahtuma ja urheilu ovat nyt omat erilliset aiheensa, ongelmallinen joulu-aihe on nyt nimetty sosiaaliseksi kanssakäynniksi, Helsinki-aihe on sulautunut tapahtuma-aiheen alle ja kesäpäivä-aihe on sulautunut urheilun alle. Perhehetki-aihe on parhaan pisteytyksen saanut ja täysin uusi aihe. Myös perhehetki-aiheen nimeäminen oli hankalaa, sillä tärkeimmät sanat eivät ole selkeästi kallellaan jonkin tietyn aihepiiriin suuntaan. Kaikki lemmatisoidut sanat sisältävän mallin tavoin tapahtuma-aihe on suomenkielisen heikoimman pisteytyksen saanut aihe.

Englanninkieliset aiheet ovat muuttuneet vähemmän, sillä vain yksi aihe on nimetty eri lailla verrattuna taulukon 18 aiheisiin: kävely-aihe on hävinnyt ja tilalle on mallintunut ruokailu-aihe, joka on huonoimman pisteytyksen saanut aihe englanninkielisessä mallissa. Urheilu-aihe on myös tässä substantiivit ja verbit huomioivassa mallissa englanninkielisen aineiston parhaimman pisteytyksen saanut aihe, tosin pisteet eivät ole yhtä hyvät kuin kaikki lemmatisoidut sanat sisältävässä mallissa. Muuten mallinnetut englanninkieliset aiheet ovat pitkälti samankaltaisia kuin taulukon 18 mallissa. Temppeliaukion kirkko on varsin spesifisesti nimetty aihe ja, kuten kaikki lemmatisoidut sanat huomioonottavassa mallissa, tässäkin mallissa se on ongelmallinen aihe, sillä se sisältää sanoja, jotka eivät suoranaisesti liity kyseiseen kirkkoon.



Kuva 52. Töölön alueen substantiivit ja verbit huomioonottavan aihemallinnuksen dominoivat aiheet aggregoituna 100 metrin kennostoon. Taustan satelliittikuva on Googlen satelliittikuva.

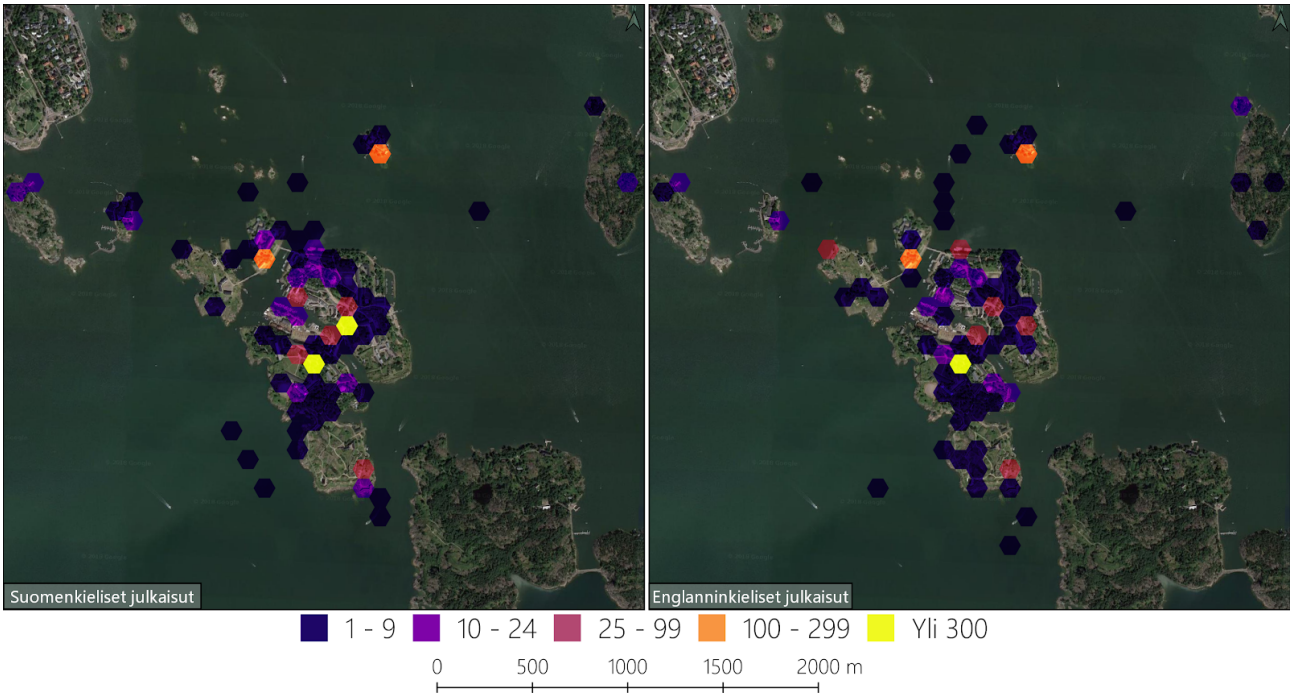
Töölön alueen aihekennostoissa kuvassa 52 on nähtävissä osittaista samankaltaisuutta erityisesti urheilu-aiheen osalta, joka muodostaa molemmissa aineistoissa selkeän yhtenäisen alueen Olympiastadionin, Kisahallin ja Helsingin jäähallin alueelle. Päällekkäisyyksiä urheilu-aiheilla on myös pienialaisemmin muualla ympäri Töölöä kuten Taivallahdessa ja Väinämöisen jalkapallokentällä. Urheilu-aiheella on molemmissa kielissä lähes yhtä suuri määrä kennoja, mutta englannin kielen urheilu-aihe voittaa muutamalla kennolla.

Suomen kielellä spatiaalisesti levinnein aihe on sosiaalinen kanssakäynti, jonka selkeimmät keskittymät ovat Taka-Töölössä Nordenskiöldinkadun/Mechelininkadun ja Mannereheimintien varrella. Muita selkeitä keskittymiä sosiaalisen kanssakäynnin aiheelle on Runeberginkadun eteläpäädyssä, Töölöntorilla, Topeliuksen puiston ympärillä ja Hesperian Esplanadin ympärillä. Valokuvaus-aihe on myös laajalle levinnyt aihe, jonka taustalla voi olla Töölön nähtävyydet, arkkitehtuuri ja itäiset sekä läntiset ranta-alueet. Valokuvaus-aiheella on laaja keskittymä Humallahdella sekä useita pienempiä kolmen aihekennon

keskittymiä niin Etu- kuin Taka-Töölössä, sekä ranta-alueilla ja myös esimerkiksi Olympiastadionin tornin kohdilla. Nämä alueet vahvistavat valokuvaus-aiheen nimeämisen onnistumista, sillä merinäkymät ja nähtävyydet ovat yleisiä valokuvauksen kohteita. Tapahtuma-aiheen spatiaalinen levinneisyys on hieman pirstaleisempaa, muodostaen yhden laajemman keskittymän Töölön eteläreunaan ja yhden hieman pienemmän pohjoisreunaan. Tapahtuma-aiheella on yksittäisiä kennoja muun muassa Oopperatalon, Kisahallin, Olympiastadionin ja Eläintarhan urheilukentän alueilla. Perhehetki on aiheista spatiaalisesti pienin levinneisyydeltään ja pienin myös julkaisumäärältään (taulukko 19). Sillä ei varsinaisesti ole keskittymiä muualla kuin Töölöntorilla Runeberginkadun varressa, jonka lisäksi aiheella on yksittäisiä dispersoituneita kennoja. Näitä ovat esimerkiksi Temppeliaukion kirkon alue, Finlandia-talo ja Urheilukadun varrella oleva Töölön pallokenttä.

Englanninkielisen mallin aiheista selkeästi spatiaalisesti laajimmalle levinnyt aihe on Suomi ja se muodostaa laajoja yhtenäisiä alueita niin Etu- kuin Taka-Töölön alueilla. Suurin yhtenäinen alue on Runeberginkadun ja Nordenskiöldin kadun välisellä alueella peittäen alleen muun muassa suuren osan Olympiastadionista, Uimastadionin ja Töölön raitiovaunuhallin kokonaan. Etu-Töölössä aiheen alla on muun muassa Finlandia-talo, Kansallismuseo ja Temppeliaukion kirkkoa ympäröivä alue, kirkko itse on oman aiheensa. Ruokailu-aihe on muuten varsin dispersoitunut spatiaalisesti, mutta sillä on muutama keskittymä esimerkiksi Nordenskiöldinkadun ja Mannerheimintien risteyksessä, Arkadiankadun varrella ja Toivo Kuulan puistossa. Useassa kennossa, joka on ruokailu-aiheen alainen, sijaitsee myös ravintoloita, joten aiheen nimeäminen on suhteellisen onnistunut. Temppeliaukion kirkon aihe on sijoittunut kartalle aivan oikeaan kohtaan, mutta kuten ylempänä mainittiin aihe pitää sisällään myös julkaisuja, jotka eivät liity kirkkoon. Tämän aiheen osalta suurin keskittymä on Runeberginkadun ja Töölönkadun risteyksessä. Aiheista spatiaalisesti pienimmälle alalle levinnyt Helsinki-aihe on spatiaaliselta rakenteeltaan varsin dispersoitunut, laajin yhtenäinen alue on Mannerheimintien varressa. Yksittäisiä kennoja on useassa paikassa ympäri Töölöä esimerkiksi Olympiastadionilla, Eduskuntatalolla ja Töölön Regattan kohdalla.

3.3.2.3 Suomenlinna



Kuva 53. Instagram-julkaisujen määrät suomen- ja englanninkielisissä aineistoissa Suomenlinnan alueelta 100 metrin kennostossa. Taustan satelliittikuva on Googlen satelliittikuva.

Suomenlinnan alueen julkaisumääriä suomen ja englannin välillä vertaillessa erot eivät ole suuria. Molemmissa aineistoissa on samoissa sijainneissa selkeät julkaisumäärien huiput. Selkeimmät erot aineistojen välillä on merialueilla. Suomenkieliset julkaisut muodostavat englanninkielisiä pidemmän yhtenäisen katkeamattoman kuvion Suomenlinnan saarella, joka yltää myös Pikku- ja Länsi-Mustan saarille. Englanninkielinen aineisto muodostaa numeroa kaksi muistuttavan kuvion, josta on erillään Kustaanmiekan sekä Pikku- ja Länsi-Mustan erilliset pienet kokonaisuudet. Hieman mielenkiintoisesti englanninkielisessä aineistossa on enemmän julkaisuja Pikku-Mustan saaressa, jossa sijaitsee Puolustusvoimien Merisotakoulu, kuin niitä on suomenkielisessä aineistossa. Alue on Puolustusvoimien hallinnoima alue, jonne ulkopuolisilla ei ole vapaata pääsyä ja tämän työn kirjoitushetkellä esimerkiksi Googlen karttapalvelussa satelliittikuva on sensuroitu Pikku-Mustan saaren osalta. Tosin usean muun karttapalvelun satelliittikuvissa saarta ei ole sensuroitu, joten kyseessä saattaa olla vain Googlen automaattinen toiminto sotilasalueille. Syynä tähän voi olla esimerkiksi Suomenlinnan saaren puolelta otetut valokuvat Merisotakoulusta, jonka sijainti myös merkitään julkaisuun, vaikka kuvan ottaminen tapahtui muualta. Toinen syy voi olla esimerkiksi se, että alueella Merisotakoulun alueella tehdyt julkaisut vain satutaan kirjoittamaan suurilta osin englanniksi.

Aineiston paikkamerkintöjä tarkastelemalla saaren paikkamerkinnät on tehty pitkälti Suomenlinnaan ja vain viidessä julkaisussa paikkamerkintä on Merisotakoulu-nimellä ja kaksi Merisotakoulun sauna -nimellä. Tämä havainnollistaa varsin hyvin työn alussa mainitun Instagram-aineiston problematiikan, mikä juontaa juurensa julkaisujen sitomisen kohdepisteisiin, eikä tarkkoihin koordinaatteihinsa. Toisaalta tämä spatiaalinen epätarkkuus on todetun etäisyydellisen vaihteluvälin sisällä (Cvetojevic et al. 2016). Suomenlinnan pohjoispuolella sijaitseva Lonnan saari vaikuttaa identtiseltä näiden aineistojen välillä. Vallisaaressa sijaitseva tumma kohta ei ole julkaisumääriä kuvaava kenno, vaan saaressa sijaitseva Vallisaaren lampi. Vallisaari oli pitkään yleisöltä suljettu saari, joka oli Puolustusvoimien omassa käytössä muun muassa sotaharjoitusalueena (Eksymä 2009). Saari avattiin yleisölle vasta vuoden 2016 kesällä, jonka vuoksi sinne merkittyjä julkaisuja tässä aineistossa ole.

Taulukko 21. Suomenkielisten Instagram-julkaisujen aiheet Suomenlinnan alueelta. Aiheiden alapuolella on tärkeysjärjestyksessä aiheiden kannalta tärkeimmät sanat. Sanojen jälkeen on aihekohtaiset koherenssipisteet, aiheeseen kuuluvien Instagram-julkaisujen lukumäärä ja aihetta kuvaava termi.

Aihe 1	**Aihe 2**	**Aihe 3**	**Aihe 4**	**Aihe 5**
suo	island	suomi	päivä	saada
kissa	käydä	linna	kesä	retki
doritsalutskijfädä	meri	kaunis	ihana	häät
kuva	ilta	helsinki	saari	lähteä
vanha	elämä	päivä	suokki	kesä
viikonloppu	pitää	retki	paras	lonta
turisti	juhla	hetki	kiva	piknikki
aurinko	poika	finland	syksy	matka
ottaa	kuva	tehdä	voida	päästä
ilta	isä	paikka	mennä	tykki
0.1321165681	**0.4406056985**	0.4124786916	0.3115320049	0.2981041495
538	322	789	628	373
Valokuvaus	**Meri**	**Suomenlinna**	**Kesäpäivä**	**Häät**

Taulukko 22. Englanninkielisten Instagram-julkaisujen aiheet Suomenlinnan alueelta. Aiheiden alapuolella on tärkeysjärjestyksessä aiheiden kannalta tärkeimmät sanat. Sanojen jälkeen on aihekohtaiset koherenssipisteet, aiheeseen kuuluvien Instagram-julkaisujen lukumäärä ja aihetta kuvaava termi.

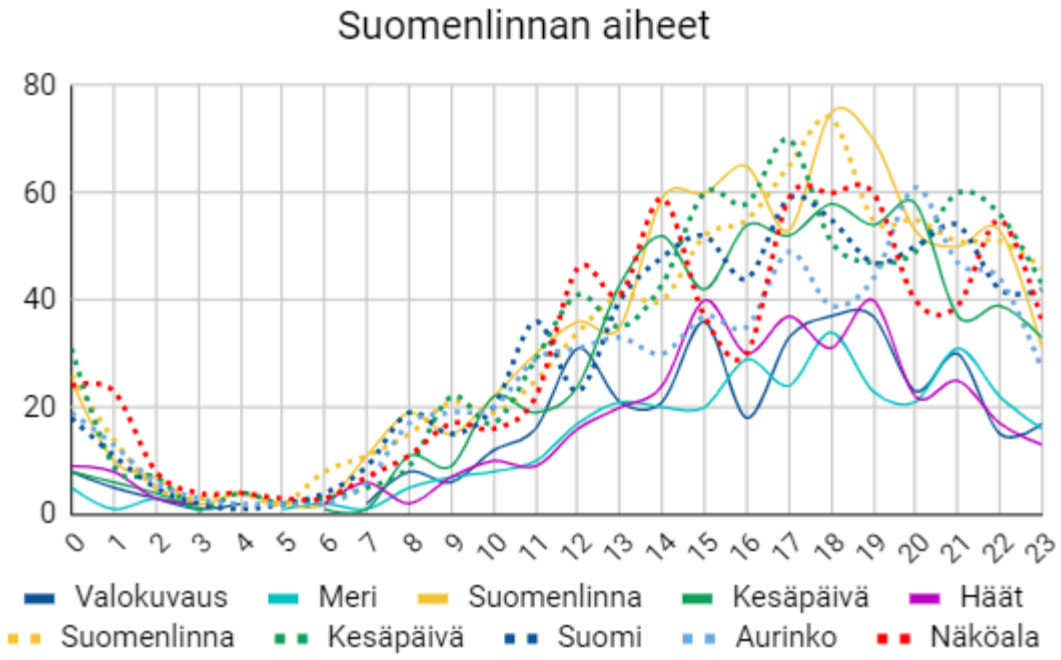
Aihe 1	**Aihe 2**	**Aihe 3**	**Aihe 4**	**Aihe 5**
fortress	helsinki	finland	day	view
island	day	beautiful	sun	will
sea	today	day	amazing	time
suomenlinna	good	suomenlinna	light	beautiful
people	summer	photo	pretty	life
helsinki	place	week	spend	good
heritage	love	cold	feel	friend
ferry	sunset	island	lovely	finnish
finland	island	picture	house	walk
site	enjoy	visit	work	year
0.5886328776	0.3609586515	0.4186150699	0.3782680225	***0.2781991959***
1015	752	698	615	704
Suomenlinna	**Kesäpäivä**	**Suomi**	**Aurinko**	**Näköala**

Taulukoihin 21 ja 22 on koostettu kaikista lemmatisoiduista sanoista muodostetut aihemallit Suomenlinnan alueelta suomen- ja englanninkielen osalta. Aiheet ovat melko samankaltaisia keskenään kielten välillä käsitellen merta, säätä ja Suomea ylipäätään molemmilla kielillä, jonka lisäksi kesäpäivä- ja Suomenlinna-aiheet löytyvät molemmista malleista. Lisäksi Suomenlinna-aihe on molemmissa aineistoissa lukumääräisesti yleisin aihe.

Koherenssipisteytyksiä tarkastelemalla suomenkielisistä aiheista parhaiten onnistui meri-aihe, joka on aluekohtaisten aihemallien osalta toiseksi parhaiten onnistunut suomenkielinen aihe. Heikoimmin onnistunut suomenkielinen aihe on valokuvaus, jossa esiintyvä omituinen sana “doritsalutskijfädä” liittyy saarella asuvaan suosittuun bloggaajaan. Suomenkielinen Suomenlinna-aihe pitää mielenkiintoisen piirteen sisällään, joka johtuu melko varmasti lemmatisointimenetelmästä. Tärkeimmät kaksi sanaa ovat suomi ja linna, jotka lienevät seurausta menetelmän tavasta käsitellä yhdyssanoja ja paikannimiä. Jostain syystä menetelmä ei ymmärtänyt Suomenlinna-sanaa paikannimenä vaan yhdyssanana, jonka menetelmä jostain syystä päätti jakaa kahdeksi omaksi sanakseen. Samankaltainen mielenkiintoinen sanan jakautuminen on tapahtunut suomenkielisessä valokuvaus-aiheen kahden tärkeimmän sanan kanssa, jossa puhekielisen Suokki-sanan inessiivi, Suokissa, on tunnistettu yhdyssanaksi, eikä Suokki-sanan inessiiviksi, ja eroteltu

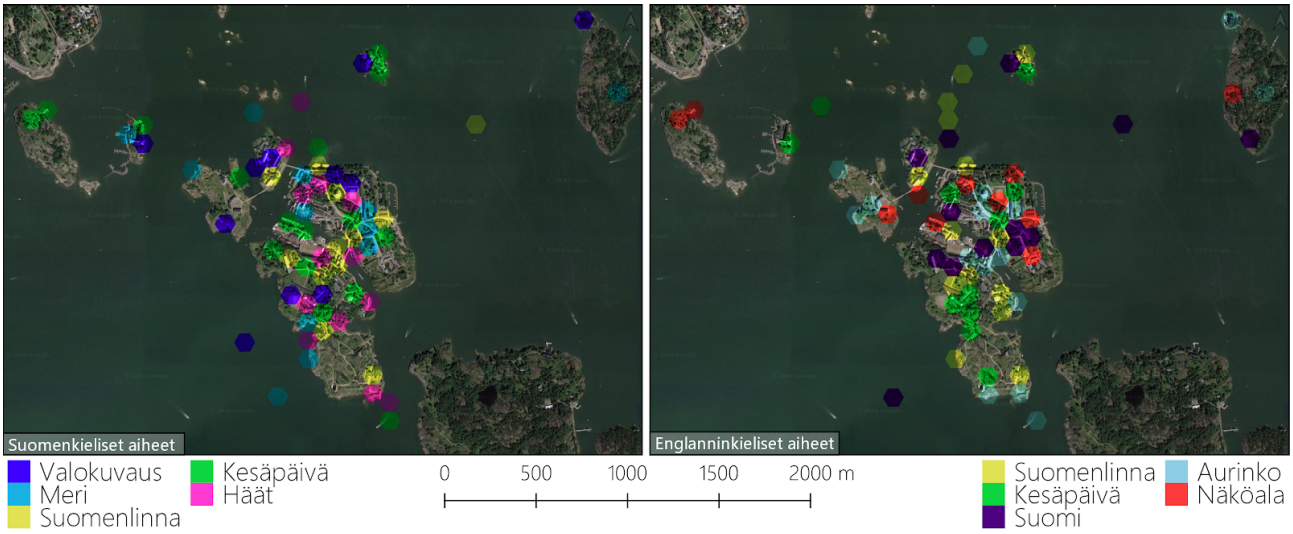
kahdeksi erilliseksi sanakseen. Puhekielisten sanojen lemmatisoimisen vaikeus ei varsinaisesti yllätä, mutta paikannimien tunnistamisen vaikeus FinnPOS-lemmatisoinnille Suomenlinnan osalta on harmillista. Tulos on varsin omituinen, sillä yhdyssanat ovat pysyneet yhdessä kaikissa aikaisemmissa aihemalleissa ja lemmatisointi on ajettu koko aineistolle identtisesti kaksi kertaa: kaikki sanat huomioivasti ja vain substantiivit, verbit ja adjektiivit huomioiden.

Englanninkielisten aiheiden osalta parhaiten onnistui Suomenlinna-aihe, joka on koko työn parhaiten onnistunut aihe koherenssipistein mitattuna. Korkeat koherenssipisteet myös näkyvät tärkeimmissä sanoissa, jotka muodostavat melko selkeän kuvan aiheesta. Heikoin englanninkielinen aihe on näköala, joka on silti tämän työn aiheiden pisteytyksiä tarkastellessa lähellä pisteiden keskitasoa oleva tulos. Maantieteellisen rajauksen pienentyessä aiheiden koherenssit tuntuvan nousevan korkeammiksi tai vähintään parhaiten onnistuneimpien aiheiden koherenssipisteet vaikuttavat olevan vakaasti melko korkeita. Englanninkielisissä aiheissa ei ole sanojen osalta samankaltaisia omituisuuksia, kuin suomenkielisen aineiston kanssa.



Kuva 54. Suomenlinnan suomen ja englanninkieliset aiheet ajassa kellonajan mukaan. Englanninkieliset aiheet on merkitty pisteviivoin ja samankaltaisilla aiheilla on samanlainen väritys.

Kuvasta 54 paljastuu Suomenlinnan temporaalinen aiherakenne, joka on selkeästi erilainen aikaisemmin havaituista temporaalisista aiherakenteista. Kolmipiikkinen rytmiikka on näkymättömissä Suomenlinnassa, englanninkielisten aiheiden julkaisumäärät ovat lähes poikkeuksetta suomenkielisiä suuremmat. Suomenlinna- ja kesäpäivä-aiheet ovat molemmilla kielillä yleisimpiä aiheita. Julkaisumäärien huippu vaikuttaisi molemmilla kielillä olevan kello 15 - 19 välillä, joka kertoo Suomenlinnan alueen luonteesta pääasiassa päiväretkikohteena molemmilla kielillä julkaiseville.



Kuva 55. Suomenlinnan kaikki lemmatisoidut sanat huomioonottavan aihemallin tulokset aggregoituna 100 metrin kennostoon. Taustan satelliittikuva on Googlen satelliittikuva.

Kuvaan 55 on visualisoitu suomen- ja englanninkieliset aiheet, jotka on mallinnettu kaikki lemmatisoidut sanat huomioonottavalla aihemallilla. Kuten kuvasta 53 jo todettiin, merialueilla ei ole kovin paljon päällekkäisyyksiä aineistojen välillä, mutta saarien osalta päällekkäisyyksiä on myös aiheiden osalta. Suomenlinna-aihe on päällekkäinen spatiaalisesti molemmissa aineistoissa esimerkiksi Pikku-Mustan eteläpäädyssä, Suomenlinnan lautan laiturin kohdalla, Susisaaren sillan ja ravintola Walhallan kohdilla. Näistä erityisesti Susisaaren silta, joka on Suomenlinnan keskiosissa, on mielenkiintoinen, sillä siellä on molemmilla kielillä julkaisujen määrältä yli 500 julkaisun kenno (kuva 53). Englanninkielisessä aineistossa Suomenlinna-aihe on laajemmalle levinnyt ja näkyy myös meren päällä, todennäköisesti lauttareitin varrella. Suomenkielinen Suomenlinna-aihe on keskittynyt pääsaarille, Iso-Mustasaareen ja Susisaareen. Kesäpäivä on myös molemmissa aihemalleisssa oleva aihe ja sen kielten välinen päällekkäisyys on vähäistä: Susisaaren keskiosissa, Suomenlinnan kirkkopuiston vieressä ja Lonnan saarella. Suomenkielinen kesäpäivä-aihe

vaikuttaa painottuvan Suomenlinnan keski- ja pohjoisosiin sekä pienille saarille, kun taas aiheen englanninkielinen vastine keskittyy Suomenlinnan eteläosiin.

Muiden suomenkielisten aiheiden spatiaalinen levinneisyys on varsin yhdenmukaista keskenään, sillä niillä on lähes yhtäläiset määrät kennoja. Valokuvaus-aihe vaikuttaa olevan keskittynyt saarten rannoille ja Suomenlinnan saaren pohjoisosiin. Meri-aiheen kennot ovat rannoilla, mutta myös Suomenlinnan Iso-Mustasaaren keskiosissa. Häät-aihe taasen on varsin tasaisesti levittäytynyt Suomenlinnan pääsaarille käsittäen Suomenlinnan kirkon ympäristön, osan Susisaaren rannikosta ja vanhan telakan. Vastaavasti muut englanninkieliset aihekennot ovat tasaisesti levittäytyneet ympäri aluetta ja aihekennojen välillä ei ole suuria lukumäärällisiä eroja. Suomi-aihe keskittyy vahvasti pääsaarille, jonka lisäksi sitä on yksittäisiä kennoja merialueella ja saarten rannoissa. Aurinko-aihe on laajimmille levinnein englanninkielen aihe ja se vaikuttaa keskittyvän Iso-Mustasaareen ja Susisaaren sillan ympäristöön, mutta myös Susisaaren eteläkärkeen, Länsi-Mustan, Lonnan ja Vasikkasaaren rannoille. Näköala-aihe puolestaan on lähes kokonaan Iso-Mustasaaressa ja kennojen sijainti antaa vaikutelman siitä, että näköala-aihe on vahva-aihe nimenomaan pohjoissuuntaisilla ranta-alueilla. Suomenlinnan pohjoispuolella on Helsingin keskusta, joka saattaa olla mieluisa valokuvattava kohde Suomenlinnasta ja lähisaarilta käsin.

Eri kieliaineistojen aiheet ovat melko samankaltaisia keskenään, kesäpäivä ja Suomenlinna ovat varsin yllätyksettömästi yhteneviä aiheita kielten välillä. Molemmilla kielillä luonnonkohteet (aurinko ja meri) ovat myös samankaltaisia aiheita keskenään. Voisikin sanoa, että molempia kieliä käyttäviä Instagram-käyttäjiä Suomenlinnaan houkuttelevat pitkälti samankaltaiset asiat, mutta selkeät eroavaisuudet antavat molemmille kieliryhmille omat piirteensä. Englanninkieliset aiheet, Suomi ja näköala, vaikuttavat keskittyvän Suomenlinnan näköalapaikkojen ja Suomen itsensä ihailuun. Tämä tulkinta saa tukea tarkastellessa molempien aiheiden osalta taulukossa 22 käytettyjä sanoja, jotka ovat suhteellisen positiivissävytteisiä, tosin Suomi-aiheessa mainitaan myös cold-sana, joka tarkoittaa kylmää. Paikkojen ja maan ihailu voi viestiä käyttäjien olevan turisteja tai kotipaikkaylpeyttä viestiviä paikallisia. Suomenkieliset aiheet, häät ja valokuvaus, kertovat Suomenlinnan olevan suosittu kohde häiden ja valokuvaamisen osalta, joista ei yhtä suoranaisesti välity vastaavanlaista kotipaikkaylpeyttä paitsi kenties välillisesti valokuvaus-aiheen osalta.

Taulukko 23. Suomenkielisten substantiivien ja verbien perusteella muodostettu viiden aiheen aihemalli Suomenlinnan alueelta.

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
kesä	kuva	päivä	saari	suomenlinna
saada	suokissa	suokki	kesäpäivä	helsinki
ilta	doritsalutskijfädä	aurinko	lonta	island
viikonloppu	tehdä	meri	häät	käydä
alkaa	ottaa	syksy	kohta	retki
lähteä	auringonlasku	mennä	sunnen	maisema
voida	pitää	paikka	seikkailla	piknikki
seura	poika	taivas	vapaapäivä	kiittää
suomi	isä	elämä	näkyä	päästä
rakkaus	kuvata	aamu	kesäretki	turisti
0.30212	0.28210	***0.19811***	**0.36162**	0.28856
414	373	389	304	840
Vapaa-aika	**Valokuvaus**	**Päiväretki**	**Häät**	**Suomenlinna**

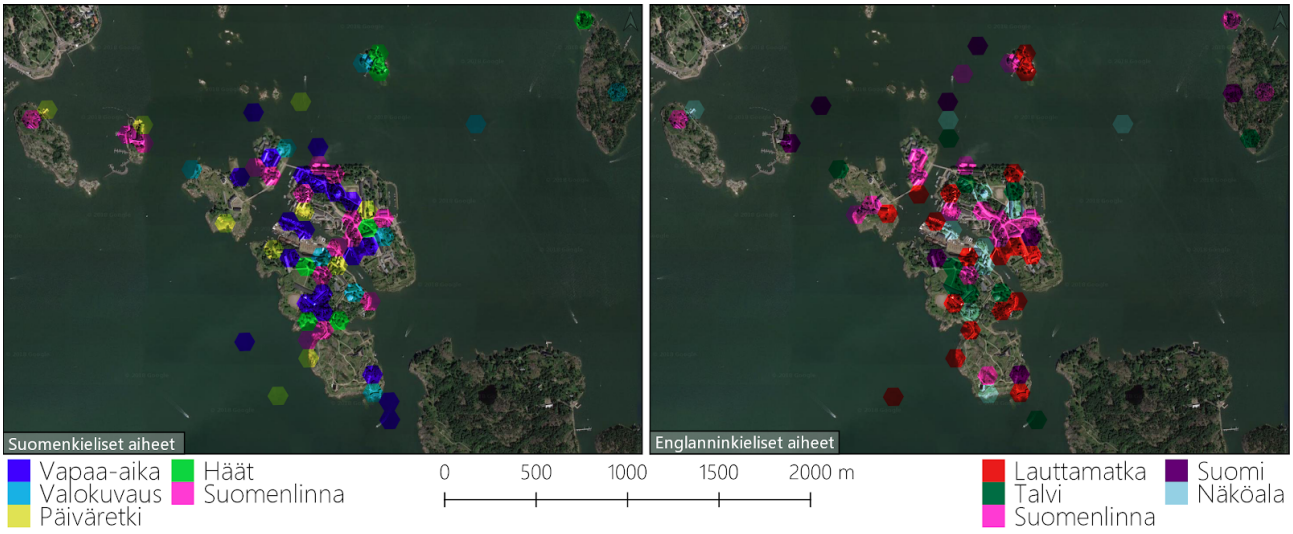
Taulukko 24. Englanninkielisten substantiivien ja verbien perusteella muodostettu viiden aiheen aihemalli Suomenlinnan alueelta.

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
helsinki	place	day	finland	sea
time	winter	island	love	view
island	people	today	walk	fortress
summer	friend	suomenlinna	find	sun
ferry	king	fortress	visit	life
will	live	photo	suomenlinna	evening
boat	house	helsinki	will	feel
city	hour	year	ice	helsinki
suomenlinna	sunset	spend	tourist	suomenlinna
weather	gate	morning	fortress	love
0.39136	0.29120	**0.48890**	***0.26910***	0.34471
680	611	852	604	658
Lauttamatka	**Talvi**	**Suomenlinna**	**Suomi**	**Näköala**

Pelkästään substantiiveja ja verbejä hyödyntävien aihemallinnuksien tulokset Suomenlinnan alueelta on koostettu taulukoihin 23 ja 24. Hieman omituisesti suomenkielisessä aihemallissa yhdyssanoja ei ole hajotettu eri sanoiksi, joka oli tapahtunut kaikki sanat huomioivassa aihemallissa. Kenties keskittyminen puhtaasti verbeihin ja substantiiveihin vaikutti lemmatisointiin FinnPOS-työkalulla eri tavoin, mutta yhdyssanat ovat pysyneet eheinä kaikissa muissa aihemalleissa taulukon 20 aihemallia lukuun ottamatta. Eri

kieliaineistojen aiheita tarkastellessa huomaa niiden muuttuneen hieman verrattuna aihemalliin, jossa käytettiin kaikkia lemmatisoituja sanoja aiheiden mallintamiseen. Suomenkieliset aiheet ovat kolmen aiheen osalta nimetty samoin ja koostuvat pitkälti samoista sanoista, mutta mallissa on näiden lisäksi kaksi uutta aihetta, vapaa-aika ja päiväretki. Näistä päiväretki oli nimeämisen osalta ongelmallinen, sillä aiheen voisi tulkita olevan myös esimerkiksi “maisema”. Lisäksi Suomenlinna-aiheen nimeämistä hieman vaikeutti puhekieliset “suokki” ja “suokissa” sanat, jotka ovat muiden aiheiden alla, mutta Suomenlinna oli viidennen aiheen tärkein sana, jonka perusteella kyseinen aihe nimettiin Suomenlinnaksi. Suomenlinna-aihe on myös selkeästi suurin aihe julkaisumääriltään. Suomenkielisistä aiheista parhaiten onnistunut on häät ja heikoimmin onnistunut on päiväretki.

Englanninkieliset aiheet ovat pysyneet myös osittain samoina, sillä taulukkoon 22 verrattuna kaksi aihetta ovat vaihtuneet taulukossa 24. Kaksi uutta englanninkielistä aihetta ovat lauttamatka ja talvi. Lauttamatka-aihe oli jälleen hieman hankala nimetä, mutta koska aiheelle tärkeimmissä sanoissa on neljä saariin ja merenkulkuun liittyvää sanaa (ferry, boat, island, suomenlinna) nimeäminen on melko onnistunut. Aiheen muut sanat ovat varsin sekalaisia teemaltaan. Talvi-aihe on varsin uusi ja nousi omaksi aiheekseen todennäköisesti adjektiivien poissulkemisen kautta, jolloin adjektiivit eivät peitä alleen tärkeitä substantiiveja ja verbejä. Parhaiten onnistunut aihe englanninkielisessä aihemallissa on Suomenlinna-aihe, joka on toiseksi yleisin aihe julkaisumääriltään. Heikoimmin onnistui neljäs aihe, Suomi, joka on myös julkaisumääriltään pienin englanninkielinen aihe.



Kuva 56. Suomenlinnan substantiivit ja verbit huomioonottavan aihemallin tulokset aggregoituna 100 metrin kennostoon. Taustan satelliittikuva on Googlen satelliittikuva.

Suomenlinnan alueelta vain substantiivien ja verbien pohjalta mallinnetut aiheet molempien kielien osalta ovat visualisoitu kartalle kuvaan 56. Ainoa kielten välinen yhteinen aihe oli Suomenlinna, jonka osalta alueellista päällekkäisyyttä on melko paljon: Iso-Mustasaaressa, Länsi-Mustassa ja Susisaaren pohjoisosissa. Selkein päällekkäisyysalue on Iso-Mustasaaressa Suomenlinnan kirjaston ja upseerikerhon välisellä alueella. Suomenlinna-aihekennot, jotka eivät ole päällekkäisiä kielten välillä, levittäytyvät eri lailla kartan muille alueille. Suomenkieliset kennot ovat levittäytyneet Särkkän ja Harakan pienille saarille, mutta myös laajemmin Susisaaren puolelle Suomenlinnaa verrattuna englanninkielisiin Suomenlinna-aihekennoihin. Englanninkieliset Suomenlinna-aihekennot ovat vastaavasti Lonnan, Vasikkasaaren ja Pikku-Mustan saarissa.

Suomenlinna-aiheen ulkopuolisten suomenkielisten aiheiden osalta spatiaalinen rakenne on varsin mielenkiintoinen. Vapaa-aika on selkeästi laajimmalle levinnyt aihe käsittäen melko laajoja yhtenäisiä alueita niin Iso-Mustasaaressa kuin Susisaaressa. Selkeimmät keskittymät ovat Suomenlinnan lautan laiturin lähistöltä upseerikerholle asti menevä alue ja Piperin kahvilan ympäristö. Valokuvaus-aihe keskittyy myös tämän aihemallin kartassa melko selkeästi ranta-alueille, kuten se teki myös kuvan 55 kartassa. Päiväretkiaihe on myös spatiaalisesti varsin rantapainoitteinen poislukien Iso-Mustasaaren Upseerikerhon viereinen alue. Esimerkiksi pienvenesatamat ja -laiturit Harakan, Särkkän, Susisaaren ja Iso-Mustasaaren saarilla ovat päiväretkiaihekennon alla. Havainto ei varsinaisesti yllätä, mutta vahvistaa nimeämisen onnistumisen todennäköisyyttä. Häät-aihe on spatiaalisesti pienimmälle alalle levinnein aihe, jonka selkein keskittymä on Lonnan saarella, jonka lisäksi yksittäisiä kennoja on Susisaaren puolella, Vasikkasaaren pohjoiskärjessä ja Sotamuseo Maneesin kohdalla.

Englanninkielen osalta lauttamatka- ja suomenlinna-aihe ovat spatiaalisesti laajimmalle levittäytyneet aiheet ja niillä on saman verran kennoja. Lauttamatka-aiheen selkein keskittymä on Iso-Mustasaaren puolella Susisaaren sillan kohdalla, jonka jälkeen pienemmät keskittymät ovat Lonnan saaressa ja sukellusvene Vesikon vieressä. Susisaaren sillan vieressä on yksityisen yhtiön maksullisen Suomenlinnan lautan laituri ja sama lautta myös pysähtyy Lonnan saaressa, joka vahvistaa kyseisen aiheen nimeämisen onnistumisen todennäköisyyttä. Talvi on varsin pienelle alalle levittäytynyt aihe, jolla on yksi selkeä keskittymä Susisaaren puolella. Muut aihekennot talvelle ovat yksittäisiä: Vasikkasaaren eteläkärki, Pikku-Mustan luoteisnurkka,

Suomenlinnan päälaiturin eteläpuolella, Kasinopuiston vieressä sekä merialueilla. Suomi-aihe on pienimmälle alalle levinnyt englanninkielinen aihe eikä se muodosta useamman vierekkäisen kennon keskittymiä lainkaan. Kolme Suomi-aihekennoa on Suomenlinnan pohjoispuolella merialueella, jonka lisäksi yksittäiset kennot Särkkän ja Vasikkasaaren saarissa. Näiden lisäksi kolme kennoa on Suomenlinnan pääsaarilla, yksi Iso-Mustasaaressa ja kaksi Susisaaressa, jonka eteläisempi kenno on Kuninkaanportin kohdalla. Näköala-aihe on myös melko pienelle alalle levittäytynyt aihe, jolla on yksi useamman vierekkäisen kennon keskittymä Susisaaren puolella Susisaaren sillan länsipuolella. Keskittymän lisäksi näköala-aihekennoja on Susisaaren eteläkärjessä, Susisaaren kapeikon vieressä, Suomenlinnan vierasvenesatamassa, Suomenlinnan kirkon vieressä ja Upseerikerhon vieressä. Muutama heikommin näkyvä näköala-aiheen kenno on merialueilla Suomenlinnan pohjoispuolella sekä Harakan saaren laiturissa.

Suomenlinnan alueella aiheet ovat kielten välillä varsin samankaltaisia keskenään, kuten asian laita on ollut myös muiden aluetasoisten aihemallien osalta, mutta selkeitä erojakin löytyy. Englanninkieliset aiheet keskittyvät enemmän paikkojen ja alueiden ihannointiin, vaikkakin se on piirre myös suomenkielisissä aiheissa. Paikkojen ihannoinnin taustalla molemmissa kieliaineistoissa lienee Suomenlinnan luonne paikkana, joka on matkailukohde suurelle osalle Instagram-käyttäjistä. Tosin suomenkieliset aiheet keskittyvät enemmän päiväretkiin, häihin ja kesäaikaan, kun taas englanninkielisissä aiheissa mainitaan näköala, kesän lisäksi talvi ja Suomi paikkana. Keskittyminen pelkästään substantiiveihin ja verbeihin toi myös Suomenlinnan alueella helppoutta aiheiden tulkintaan ja nimeämiseen.

3.3.2.4 Adjektiivimallinnus

Lemmatisoituja substantiiveja ja verbejä hyödyntäneiden aluekohtaisten aihemallien lisäksi myös pelkästään adjektiiveihin kohdistunut aihemallinnus toteutettiin aluekohtaisesti. Adjektiiviaihemallinnuksen toteuttamisen taustalla oli idea kokeilla soveltuvatko mallinnuksen tulokset kevyeen sävyanalyysiin, jota pystyisi hyödyntämään digitaalisen kaupunkitilan kuvailussa ja sen syventämisessä. Adjektiiveja käyttävät alueelliset aihemallit onnistuivat teknisesti ottaen varsin hyvin, mutta syntyneet tulokset ovat lähes käyttökelvottomia eivätkä valitettavasti tuo lisää syvyyttä tuloksiin. Jo ennen adjektiivimallinnuksen toteuttamista, aineistoon syventymisen yhteydessä, adjektiivien käyttäminen aihemallinnuksessa eräänlaisena alkeellisena sävyanalyysina alkoi vaikuttaa huonolta idealta. Aineistossa esiintyvät adjektiivit

ovat vahvasti positiivissävytteisiä, kuten huomattiin jo lemmatisoiduista sanoista muodostetuissa sanapilvistä kuvissa 33 ja 34. Sävyerojen puuttuessa tai ollessa erittäin vähäisiä, mallinnettavien aiheiden välisten merkityksellisten erojen eritteleminen hankaloituu huomattavasi. Semanttisesti merkityksellisten erojen löytäminen erittäin samankaltaisten adjektiivien välillä on vaikeaa, eikä erojen kvantifioiminen esimerkiksi pisteytyksen keinoin vaikuta kovin järkevältä, saati työn tarkoitusperiä tukevalta tehtävältä. Esimerkkinä adjektiiviaihemallien hyödyntämisen vaikeudesta taulukkoon 25 on koostettu suomenkielisten adjektiivien aihemalli Töölön alueelta, josta on helposti havaittavissa sävyjen erottelun ja pisteyttämisen hankaluus. Esimerkiksi onko “kaunis” positiivisempi sana kuin “paras”, “ihana”, “kiva” tai Näiden seikkojen vuoksi, adjektiivipohjaisia aihemalleja ei tässä työssä käytetä Helsingin digitaalisen kaupunkitilan tutkimiseen ja erittelemiseen, sillä ne eivät tuo analyysiin toivottua lisäsyvyyttä. Lisäksi, sävyanalyysia ei tulisi tehdä aihemallien avulla vaan sitä varten kehitetyillä työkaluilla. Valitettavasti suomenkielelle tehtyjä työkaluja ei ollut tämän työn kirjoittamishetkellä avoimesti saatavilla, eikä kirjoittajan osaamistaso riitä tämänlaisen työkalun luomiseen itsenäisesti. Englannin kielelle työkaluja ja valmiita sävymalleja olisi saatavilla, mutta koska sävyanalyysi on vahvasti kontekstiriippuvaista (Taboada et al. 2011), vapaasti saatavilla olevat sävymallit olisivat soveltuneet huonosti tämän aineiston sävyjen analysointiin. Esimerkiksi sävymalli, joka on rakennettu esimerkiksi jonkin kaunokirjallisuuden teoksen sanastolla, sopii huonosti sosiaalisen median aineiston sävyanalyysiin. Adjektiiveihin perustuvan aihemallinnuksen ei hyödynnetä digitaalista kaupunkitilaa eriteltäessä tässä kappaleessa mainituista syistä johtuen.

Taulukko 25. Töölön alueelta muodostettu aihemalli, jossa on hyödynnetty vain Töölön suomenkielisten Instagram-julkaisujen kuvatekstien adjektiiveja.

Aihe 1	**Aihe 2**	**Aihe 3**	**Aihe 4**	**Aihe 5**
kaunis	paras	ihana	kiva	eka
valmis	upea	mahtava	iso	pitkä
vähä	hieno	rakas	kova	hauska
parka	suomalainen	tärkeä	eilinen	huikea
vanha	puhdas	iloinen	mukava	onnellinen
tällainen	oikea	täysi	nuori	täydellinen
pikku	lämmin	sininen	mieletön	erilainen
parhain	aurinkoinen	jännä	huono	punainen
seuraava	perinteinen	illallinen	loistava	siisti
huominen	kuuma	reilu	herkullinen	ilmainen
0.37402	0.26307	*0.1469*	0.247	0.2795
12976	1385	1557	1170	1204

4.0 Keskustelu

4.1 Julkaisujen aiheet liittyvät vapaa-aikaan

Analyysi osoittaa, että suomen- ja englanninkielisten Instagram-julkaisujen aiheet liittyvät vahvasti vapaa-aikaan. Vahva kytkös vapaa-aikaan paljastui jo ennen aihemallinnusta, kun käsittelemätöntä aineistoa visualisoitiin ajassa. Yleisimmät kellonajat julkaista Instagramissa ovat yleiset ruokailuajat, jonka lisäksi viikonpäivittäin tarkasteltuna viikonloppu on selkeästi suosituin aika viikosta julkaista. Samanlaisia vapaa-aikaan viittaavia ajallisia piirteitä paljastui myös aiheiden temporaalista jakautumista tarkastellessa, tosin aiheiden ja kielten välillä oli hieman eroja. Kytkös vapaa-aikaan vahvistuu, kun suomen- ja englanninkielisten Instagram-julkaisujen kuvatekstien aiheet mallinnettiin ensiksi Helsingin tasolla ja sen jälkeen kaupunginosatasolla kolmen kaupunginosan julkaisujen kautta. Lähes jokaisessa aihemallissa kielestä riippumatta on ruokaan, sosiaaliseen kanssakäymiseen, juhlimiseen ja juhlapyhiin liittyviä aiheita. Tosin yleisestä vapaa-aikaan liittyvästä aihemassasta poikkeavia aiheita löytyy ja tämä näyttäytyy erityisesti kielten välillä. Kielten väliset aihe-erot eivät ole semanttisesti kovin suuria tai selkeästi vastakkaisia, vaikka vastakohtaisuutta esiintyy hieman kielten välillä sekä myös kielen omien aiheiden keskuudessa. Vastakohtaisuus on melko voimakas ilmaisu, sillä aineiston aiheissa ei ilmene vahvaa polarisaatiota kieliryhmien välillä eikä sisällä, vaan vastakohtaisuus ilmenee lähinnä hieman selkeämmin “harmaaseen arkeen” ja hieman selkeämmin vapaa-aikaan liittyvien aiheiden kautta. Aiheet muuttuvat tarkastelualueen ja kielen vaihtuessa jonkin verran, mutta muutamat aiheet esiintyvät lähes kaikilla tarkastelutasoilla, kuten ruokailu ja sosiaalinen kanssakäynti. Muutamilla alueilla kieltenväliset aihe-erot ovat selkeämpiä kuten Töölössä, mutta joillain alueilla erot ovat hienovaraisempia kuten Suomenlinnassa.

Kymmenen aiheen aihemalleissa aiheiden vastakohtaisuutta on suomenkielisten aiheiden keskuudessa, mutta myös kielten välillä. Työ ja arkielämä ovat suomenkielisistä aiheista selkeimmin erillisiksi erottuvat aiheet, sillä muut aiheet ovat vahvasti vapaa-aikaan liittyviä. Nämä kaksi aihetta ovat Jan Gehlin aktiviteettitaulukon (taulukko 1) mukaisesti “pakollisia” aktiviteetteja. Sanasisällöiltään ne ovat myös varsin neutraaleja aiheita verrattuna muihin aiheisiin. Englanninkielisissä aiheissa ei näy vastaavaa “harmaan arjen” ja vapaa-ajan välistä vastakkainasettelua. Suomenkielisten julkaisujen aiheiden näkyvä kytkös

arkiseen elämään oli yksi ennakko-odotuksista aineistoa kohtaan, joka vaikuttaa toteutuneen siitä huolimatta, että molempien kielien Instagram-aineiston todettiin olevan selkeästi kytköksissä vapaa-aikaan niin temporaaliselta, spatiaaliselta kuin aihe-rakenteeltaan. Aiheet, kuten mainitut työ ja arkielämä, kytkevät suomenkielisen aineiston hieman vahvemmin arkiseen elämään, jonka lisäksi englanninkielisessä aineistossa nähdyt erityisiin paikkoihin liittyvät aiheet eivät näkyneet lainkaan suomenkielisissä aiheissa Suomenlinnaa lukuun ottamatta. Tämä kertoo siitä, että suomenkielisille Instagram-käyttäjille nähtävyydet ovat vain osa arkiympäristöä, eivätkä ne siten ole houkuttelevampi Instagram-julkaisun aihe kuin mikään muukaan arkisen fyysisen ympäristön kohde. Havainto heijastaa kirjallisuudessa mainittuja käyttäytymismalleja paikallisten ja matkailijoiden välillä (Urry & Larsen 2011, 13-18), mutta myös ylipäätään eri kieliryhmien välisiä eroja (Graham & Zook 2013). Näistä arkeen liittyvistä piirteistä huolimatta suomenkieliset aiheet ovat myös selkeästi vahvemmin kytkeytyneet vapaa-aikaan kuin arkeen.

Vuodenajat näkyvät vastakkaisina kielten välisissä aiheissa: suomenkielisissä aiheissa pinnalle nousee useammin kesä, kun taas englanninkielisissä aiheissa talvi. Aikaan liittyviä mielenkiintoisia piirteitä on myös aiheiden sisällä, sillä jotkin aiheista pitävät sisällään anakronistisia sanoja: esimerkiksi joulu-aihe voi pitää sisällään sanan kesä. Anakronistiset tai muuten aiheestaan “poikkeavat” sanat aiheuttivat vaikeuksia aiheiden nimeämisessä muutaman aihemallin kohdalla, mutta aiheiden alueellisten ja ajallisten tarkastelujen kautta sanastoltaan ristiriitaisiin aiheisiin saatiin lisää selkeyttä. Vastakohtaiset sanat aiheiden sisällä eivät kerro aiheen epäonnistumisesta vaan ne pikemminkin heijastavat monimuotoisten kuvatekstien sisältöä. Esimerkiksi joulu-aiheeseen sisältyvä kesä-sana voi hyvinkin johtua useasta kuvatekstistä, jossa saatetaan toivottaa hyvää joulua, jonka perään toivotaan kesän pikaista saapumista. Aiheiden sisäiset ristiriitaiset ja vastakohtaiset sanat heijastavat ihmisten ilmaisunvapautta, joka mahdollistaa Rosen (2017) kuvaileman monimuotoisten identiteettien muovaaman digitaalisen kaupunkitilan syntymisen.

Englanninkieliset aiheet koko Helsingin alueelta heijastelevat pitkälti vapaa-ajan aktiviteetteja kuten suomenkielisetkin aiheet, mutta niissä on nähtävissä paikkoihin liittyvien aiheiden kautta englanniksi julkaisevien turistien vaikutus ja mainitun “arkisuuden” näkymättömyys. Esimerkiksi erilaisten paikkojen, alueiden ja suomeen liittyvien aiheiden näkyminen useissa englanninkielisissä aihemalleissa, mutta niiden

puuttuminen suomenkielisistä, voi olla merkki johdantokappaleessa mainitusta turistien taipumuksesta tarkastella matkakohteensa “eksoottista” ympäristöä eri tavoin kuin arkiympäristöään (Urry & Larsen 2011, 13-18). Suomeksi julkaiseville esimerkiksi Temppeliaukion kirkko ei ole paikka, johon tavalliset arkiset julkaisut liittyvät, vaan se nousee esiin vasta esimerkiksi erityisen tapahtuman kuten häiden tai joulujuhlien kautta, mutta silloinkaan itse kirkko ei välttämättä ole julkaisun aihe vaan jokin tapahtuma Tosin myös suomalaiset julkaisevat englannin kielellä, joten englanninkielisissä aiheissa näkyvät alueet, paikat ja suomalaisuutta käsittelevät aiheet voivat osittain liittyä “kotipaikkaylpeyteen” tai kuuluvuuden tunteeseen, jota Instagramin käyttäjät haluavat viestiä kansainvälisesti käytetyllä kielellä. Tämän piirteen vuoksi englanninkielistä aineistoa ei voi suoranaisesti pitää erityisen puhtaana “turistiaineistona”, vaikkakin sen suuntaisia piirteitä sillä on runsaasti (ks. myös Hiippala et al. 2018).

Aluekohtaisista tarkasteluista näkyy koko Helsingin mittakaavan aihemalleissa todetut erot: suomenkielisissä aiheissa on usean vapaa-aikaan liittyvän aiheen lisäksi usein muutama arkisempi aihe. Tosin Kallion alueella englannin kielisissä aiheissa arkeen liittyvä aihe nousee näkyviin. Töölö on kieltenvälisten aihe-erojen kautta selkeästi jakautunein alue, Suomenlinna yhtenäisin ja Kallio näiden välissä, tosin lähempänä Suomenlinnaa. Kielten väliltä löytyy myös hieman hienovaraisempia eroja, joista voi huomata suomenkielisten julkaisujen kytköksen arjen lisäksi suomalaiseen kulttuuriin, esimerkiksi häät-, joulu-, pikkujoulut- ja kesä-aiheiden kautta. Esimerkiksi joulu- ja kesä-aihe ovat melkein jokaisessa suomenkielisessä aihemallissa, mutta eivät juuri lainkaan englanninkielisissä malleissa. Englanninkieliset aiheet keskittyvät vahvemmin paikkoihin, talveen ja yleispiirteisesti sosiaalisiin vapaa-aikaan liittyviin aiheisiin, kuten juhlintaan ja ystäviin. Aihe-erojen osalta tuloksista voidaan lyhyesti sanoa englanninkielisten aiheiden keskittyvän vahvemmin paikkoihin ja suomenkielisten aiheiden keskittyvän vahvemmin arkeen ja tapahtumiin. Kaupunginosien välisten erojen näkyminen ei ole yllättävää, sillä eri kaupunginosiin sidottujen sosiaalisen median julkaisujen on todettu eroavan toisistaan myös aikaisemmissa tutkimuksissa (Lansley & Longley 2016; Martin & Schuurman 2017; Fu et al. 2018).

Kielten välisistä eroista huolimatta aihekokonaisuudet olivat melko samankaltaisia kussakin tarkastelussa. Aiheiden samankaltaisuutta voi selittää pitkälti Instagram-alustan profiloituminen vahvasti positiivisiin

asioihin painottuvalle sisällölle (Keipi et al. 2018). Helsingin suomen- ja englanninkieliset Instagram-julkaisut ovat osoittautuneet tapahtuvan yleisinä ruoka-aikoina, erityisesti viikonloppuisin, läheltä palveluja ja koostuvan pitkälti positiiviseen sävyyn kirjoitetusta sisällöstä. Tästä syystä osa kielten välisistä aiheista muistuttavat toisiaan tai ovat miltei identtisiä keskenään niin koko Helsingin mittakaavassa, kuin myös kaupunginosatason tarkastelussa. Aiheiden nimeäminen on tämän työn haastavimpia vaiheita johtuen subjektiivisuudestaan, mutta siihen tukeuduttiin kirjallisuudessa käytettyjen esimerkkien mukaisesti ja nimillä lähinnä pyritään helpottamaan aiheiden erottamista toisistaan.

Aihemallinnuksen tuloksia voisi parantaa esimerkiksi sillä, että yleisesti tunnetut käsitteet kuten “uusi vuosi” säilyisivät esikäsittelyssä ehjänä, jolloin aihemalli käsittelisi niitä yhteen kuuluvana käsitteenä useamman erillisen sanan sijaan. Tähän voisi sopia n-grammiin perustuva vektorirepresentaatio, jossa sanaa edeltävä ja sitä seuraava sana huomioidaan (Esposito et al. 2016; Arslan et al. 2018). Näissä tuloksissa uuteen vuoteen liittyneet julkaisut eivät erottuneet kovin selkeästi, sillä sanat uusi ja new ovat hukkasanoja, joten ne jätettiin huomioimatta aihemallissa. Tuloksiin voisi saada myös lisää voimakkuutta, mikäli erikielisten aihekokonaisuuksien samankaltaisuutta voisi vertailla tilastollisesti. Nyt aihe-erot ovat pitkälti subjektiivisen tulkinnan varassa, vaikkakin aiheiden nimeämistä tuettiin tarkastelemalla muodostuneita aiheita ajassa ja paikassa.

Jatkoa ajatellen suomalaisten englanniksi tekemät Instagram-julkaisut voisivat soveltua erinomaisesti esimerkiksi useaan otteeseen mainitun kotipaikkaylpeyden, kuulumisen tunteen ja kielenvalinnan taustatekijöiden tutkimiseen. Mikä saa pääsääntöisesti suomeksi julkaisuja tekevän käyttäjän kirjoittamaan julkaisun englanniksi ja mitä tekijöitä julkaisun kielen valitsemisen taustalla on? Tapahtuuko kielenvalinta tietyn aiheen, ajanjakson tai paikan ympärillä? Julkaisevatko suomalaiset enemmän suomeksi esimerkiksi arkisista asioista, mutta julkaisut arjesta poikkeavista tapahtumista kuten lomamatkoista tai juhlista tehdään englanniksi? Kielenvalintaan vaikuttaneita muuttujista ei voi tämän työn puitteissa sanoa mitään varmaa, mutta mahdollisia vaikuttajia voi pohtia. Esimerkiksi valinta käyttää englantia suomen sijasta saattaa johtua siitä, että julkaisuun liittyy jokin asia, jolle käyttäjä haluaa kansainvälistä näkyvyyttä kuten lomamatka, juhla tai muu hänelle tärkeä asia. Kansainvälisen näkyvyyden vahvistaminen voi juontaa juurensa haluun julkaista

kielellä, joka on kyseisen käyttäjän suomea osaamattomien ystävien tai seuraajien ymmärrettävissä. Se voi myös tarkoittaa sitä, että käyttäjä haluaa julkaisunsa tavoittavan mahdollisimman monta muuta käyttäjää ja muuttuvan "viraaliksi", jolloin käyttäjälle täysin tuntemattomat muut käyttäjät innostuvat julkaisusta ja levittävät sitä omiin verkostoihinsa. Hän voi myös yksinkertaisesti haluta ylläpitää kielitaitoaan.

4.2 Aiheiden spatio-temporaalinen rakenne on kaksijakoinen

Kielten väliset aihe-erot ja samankaltaisten aiheiden erilainen sijoittuminen ajassa ja paikassa kertovat kielten välisistä eroista digitaalisessa kaupunkitilassa. Ennen aiheisiin menoa, jo maankäyttöluokkiin perustunut julkaisumäärien vertailu (kuva 31) paljasti kielien välillä olevan alueellisia eroja, vaikkakin sijoittumislogiikka on pääpiirteittäin samankaltainen. Mikään yksittäinen maankäyttöluokka ei ollut erityisen vahvasti toisen kielen dominoima, vaikkakin lähes kaikissa maankäyttöluokissa suhteellisia eroja oli nähtävissä. Maankäytön mukaan luokitellut suomen- ja englanninkieliset julkaisut tapahtuvat pääasiallisesti tiiviin rakentamisen alueilta kuten palvelualueilta, tiiviiltä asuinalueilta ja liikennealueilta. Näiden alueiden ulkopuolella suomeksi kirjoitettuja julkaisuja on vahvemmin tavalliseen arkeen ja työskentelyyn liittyvissä maankäyttöluokissa kuten rakennustyömailla ja maatalousalueilla, kun taas englanniksi kirjoitetut julkaisut olivat vahvemmin esillä satama- ja lentokenttäalueilla, joka kertonee turismin vaikutuksesta.

Kielidominanssi paljasti englannin ja suomen kielten jakavan Helsingin keskustan selkeästi kahtia, englanninkielisten julkaisujen hallitessa eteläpuoliskoa ja suomenkielisten hallitessa alueita Hakaniemestä pohjoiseen. Aikatarkastelussa tunnistettujen kielten osalta suurin vaihtelu oli nähtävissä muiden kielien määrissä, mutta suomen ja englannin aseman säilyminen toisiinsa ja muihin kieliin nähden läpi aineiston on melko kiinnostava havainto. Se voi viestiä suomen- ja englanninkielisten käyttäjien yhtäläisestä määrän kasvusta Instagramissa Helsingin alueella, joista osa lähes varmasti on englanniksi julkaisevia suomalaisia käyttäjiä. Suomen ja englannin kielillä oli samankaltainen temporaalinen kolmipiikkinen vuorokausirytmiikka ja viikonloppuvoittoinen julkaisumäärähuippu. Pelkästään suomen- ja englanninkielisten julkaisujen spatio-temporaalinen rakenne viittaa digitaalisen kaupunkitilan erilaisuuteen kielten välillä: molemmat ovat ajallisesti kytköksissä vapaa-aikaan, mutta suomen- ja englanninkieliset julkaisut hallitsevat eri alueita

kaupungissa ja maankäyttöön perustuvan vertailun perusteella suomenkieliset julkaisut ovat vahvemmin kytköksissä arkielämään ja englanninkieliset julkaisut puolestaan turismiin.

4.2.1 Aiheiden erot ajassa

Aiheiden ajallinen tarkastelu koko Helsingin mittakaavassa paljasti, että aiheilla on toisiaan pitkälti vastaavat rytmiikat vuorokaudenaikojen ja viikonpäivien mukaan. Rytmiikka ei välttämättä ole kytköksissä Instagram-käyttäjiin ja heidän toimintatapoihin, sillä samankaltaista julkaisurytmiikkaa on paljastunut myös Suomi24 -keskustelufoorumilla (Lagus et al. 2016). Yleisimmät ruokailuajat näyttäytyivät selkeinä piikkeinä aiheiden osalta molemmilla kielillä vuorokaudenaikoihin perustuvassa vertailussa, mutta muuten kielten välillä oli selkeitä eroja.

Suomenkieliset aiheet seurasivat toisiaan melko lähekkäin vuorokausirytmissä: esimerkiksi mikään aihe ei selkeästi pudonnut tai kasvanut muihin aiheisiin nähden vaikkakin suomenkielisten aiheiden keskinen hierarkia vaihteli. Arkisten asioiden yleisyys suomenkielisten aiheiden tarkastelussa kellonajan mukaan vahvistaa kyseisen aineiston kytköstä entisestään arkiseen elämään: aamut alkavat aiheissa urheilulla ja aloittamisella, keskipäivät täyttyvät arkielämästä ja illalla sosialisoidaan ja urheillaan jälleen. Vasta viikonpäivittäinen tarkastelu tuo suomenkielisiin aiheisiin selkeää aiheiden välistä vaihtelua: arkielämä on arkena pitkälti yleisin aihe, mutta viikonlopun tullen urheilu ja sosiaalinen kanssakäynti ottavat selkeän enemmistön julkaisujen aiheista. Sosiaalinen kanssakäynti säilyttää korkean asemansa myös sunnuntaina, jolloin kaikki muut julkaisujen aiheet putoavat yleisyydessä. Tämä kertonee suomeksi julkaisevien Instagram-käyttäjien viettävän paljon aikaa ystävien tai perheiden seurassa lauantaisin ja sunnuntaisin. Kuukausiperusteinen suomenkielisten aiheiden tarkastelu paljasti Instagram-käyttäjämäärän kasvun, vahvisti osan suomenkielisten aiheiden nimien nimeämisen onnistumista sekä muiden aikatarkastelujen havaintojen perusteella tehtyjä johtopäätöksiä. Kaiken kaikkiaan suomen kielen osalta Instagram-julkaisujen aiheet vaikuttavat heijastavan käyttäjiensä arkea melko hyvin.

Englanninkielisten julkaisujen osalta aikatarkasteluissa on suomeen verrattuna selkeitä eroja. Aiheiden tarkastelu kellonajan mukaan paljastaa englanninkielisillä aiheilla olevan sama kolmen julkaisupiikin rakenne

kuin suomeksi kirjoitettujen julkaisujen aiheilla, mutta aiheiden väliset erot julkaisumäärissä ovat selkeästi erilaiset. Aamu-aihe hallitsee erityisesti aamupäivää, mutta pysyy yleisyydessään korkealla läpi koko päivän. Kenties Helsingin aamu on niin erityinen tapahtuma englantia käyttäville Instagram-käyttäjille, että se ansaitsee oman julkaisunsa. Toisaalta useasti mainittu mahdollinen matkailun vaikutus voi nostaa aihetta, sillä Helsinki on heille eksoottinen paikka, jolloin myös helsinkiläinen aamu voi olla heille erityinen kokemus. Ruokailu ja uusivuosi ovat myös muiden aiheiden kellonajallisesta yleisyydestä poikkeavia aiheita suuremmalla yleisyydellään. Aiheen nimeäminen uudeksivuodeksi oli hankala päätös, mutta kuten kappaleessa 3.3.1 käytiin lävitse, se lienee silti kuvaavin nimi kyseiselle aiheelle, vaikka esiintyykin yleisenä aiheena kellon- tai vuodenajasta riippumatta.

Viikonpäiväperusteisessa tarkastelussa englanninkielisessä aineistossa on suomenkielisestä poikkeava rakenne. Lauantai on edelleen suurin julkaisupäivä, mutta arkipäivien välillä on selkeää hajontaa aiheiden yleisyydessä eivätkä kaikki aiheet kasva kohti viikonloppua. Usean aiheen yleisyys aaltoilee viikon aikana, jonka taustalla voi olla turistien ja matkailijoiden vaikutus, sillä lomailijat eivät todennäköisesti välitä kovin paljoa kellonajoista ja siitä onko arkipäivä vai viikonloppu. Viikonloppuun liittyen, aiheista viikonloppu on nimetty erityisen onnistuneesti, sillä se on arkipäivinä harvinaisin aihe, mutta kasvaa yleisyydeltään miltei kaksinkertaiseksi viikonloppuisin. Usea englanninkielinen aihe myös jostain syystä harvinaistuu perjantaisin. Taustalla voi olla viikonloppuaiheen räjähdysmäinen kasvu perjantaina tai jokin muu vaikuttaja. Toisin kuin suomenkielisissä aihemalleissa melko useat englanninkieliset aiheet kasvavat vielä lauantaista sunnuntaihin yleisyydellään, jonka taustalla saattanee olla useaan otteeseen mainitut turistit ja heidän vapaus arjen rytmityksestä. Englanninkieliset aiheet tuntuvat heijastavan arkista elämää aikatarkastelun perusteella selkeästi vähemmän suomenkielisiin aiheisiin verraten, joten on mahdollista, että ne kuvaavat ainakin epäsuoralla tasolla turistien näkemää digitaalista kaupunkitilaa.

On selvää, että suomen- ja englanninkielisten Instagram-julkaisujen aiheissa on eroja ajassa nähtynä, vaikkakin yleispiirteinen rytmiikka on samankaltainen kielten välillä koko Helsingin mittakaavassa. Kaupunginosatasolla kolmipiikkinen rytmiikka ei ollut aivan yhtä selkeästi näkyvissä: näkyvin se oli Töölössä, heikosti näkyvä Kalliossa ja käytännössä näkymätön Suomenlinnassa, tosin näissä alueellisissa julkaisurytmin

muutoksissa ei ollut kielten välillä kovin suuria eroja. Suomenlinnan aikarakenne heijastaa alueen samankaltaisuutta suomeksi ja englanniksi julkaiseville Instagram-käyttäjille. Ylipäätään suomenkielisten aiheiden aikarakenteessa näkyy arjen vaikutus, kun taas englanninkielisten aiheiden aikarakenteessa näkyy mahdollisesti matkustajien vaikutus.

4.2.2 Aiheiden erot alueellisesti

Aiheiden spatiaalinen tarkastelu koko Helsingin mittakaavassa paljasti spatiaalisen aiherakenteen monimuotoisen ja vaikeaselkoisen luonteen. Aiheet eivät muodosta kovin selkeitä suuria kokonaisuuksia ja niiden sijoittuminen vaikuttaa osittain satunnaiselta. Samannimisten ja samankaltaisten aiheiden osalta Helsingin mittakaavassa oli eroja aihesolujen lukumäärissä, mutta samanlaisten aiheiden spatiaalinen päällekkäisyys vaikutti enimmäkseen olevan satunnaista ja tapahtuvan vain muutamien solujen ryppäille kerrallaan. Toisaalta aiheen ja sijainnin välinen yhteys vaikutti olevan ainakin osittain looginen, esimerkiksi urheilu-aihe oli yleisin aihe yleisten urheilupaikkojen läheisyydessä kuten Olympiastadionilla, Helsingin jäähallilla ja Uimastadionilla. Instagram-julkaisujen aiheiden ei voi olettaa liittyvän aina siihen sijaintiin, mihin se on liitetty (Cvetojevic et al. 2016; Martin & Schuurman 2017; Fu et al. 2018), joten sekava spatiaalinen rakenne ei suoranaisesti osoita aineiston laadun olevan huono tai analyysin epäonnistuneen, vaan heijastavan suurten ihmisjoukkojen monimuotoisuutta. Monimuotoisuus tosin vaikeuttaa aiherakenteen havainnollistamista ja visualisointia etenkin laajoilla alueilla, jonka vuoksi Instagram-julkaisujen kuvatekstien aiheiden tarkastelu eri paikallisilla aluetasoilla voi tehdä siitä havainnollisempaa (Jenkins et al. 2016). Lisäksi tällöin on nähtävissä, toistuuko samankaltainen aiheiden spatiaalinen rakenne, yleispiirteisen tason lisäksi myös pienemmässä mittakaavassa.

Kaupunginosatasolla tapahtunut aihetarkastelu paljasti, että aiheiden spatiaalinen eroavaisuus ja osittainen päällekkäisyys siirtyvät yleisemmältä tasolta myös tarkemmalle tarkastelutasolle. Tarkempi kaupunginosatason tarkastelu myös paljasti Instagram-aineiston paikallistason aiherakenteen olevan vahvemmin kytketty fyysiseen ympäristöönsä, sillä aluetason tarkasteluissa kyseisten kaupunginosien nähtävyydet ja erityispiirteet alkoivat näkyä aiheiden tärkeimmissä sanoissa. Digitaalisen ja fyysisen ympäristön on todettu heijastavan toisiaan myös muualla (Tranos & Nijkamp 2013). Lisäksi aihemallien

koostaminen uudestaan pelkästään substantiiveja ja verbejä käyttämällä vaikutti terävöittävän aiheita entisestään.

Samannimisiä aiheita on molemmilla kielillä kaupunginosatasolla, yleisimpinä ruokailu, joulu ja sosiaalinen kanssakäynti. Samannimisten aiheiden aihekennot sijoittuvat muutamista päällekkäisyyksistä huolimatta pääasiallisesti eri alueille, kuten koko Helsingin mittakaavan tarkastelussakin. Havainto heijastaa alueellista "kielellistä epätasaisuutta", jonka myötä eri alueet näyttäytyvät eri kieliryhmille eri tavoin Internetin ja älylaitteiden kautta tarkasteltuna (Graham & Zook 2013). Esimerkiksi ruokailun ja sosiaalisen kanssakäynnin aiheet olivat kaupunginosasta riippumatta pitkälti eri alueilla kielten välillä. Kenties eri ravintolat houkuttelevat englanniksi julkaisevia ihmisiä, kuin ravintolat joista tehdään suomeksi julkaisuja. Tämä ei vaikuta yllättävältä ottaen huomioon tässä työssä paljastuneen englanninkielisen aineiston kytkeytymisen turismiin. Matkailijat todennäköisesti haluavat kokeilla suomalaisia tai pohjoismaisia erikoisruokia enemmän kuin suomalaiset itse tai vähintään he tekevät niistä mieluummin julkaisun sosiaaliseen mediaan. Muutama aihe, kuten Töölössä esiintynyt urheilu, oli melko selkeästi päällekkäinen spatiaalisesti kielten välillä, sillä Olympiastadionin, Helsingin jäähallin ja Kisahallin muodostamalla alueella tapahtuu paljon urheiluun liittyviä asioita, eikä samankaltaisia urheilupaikkoja ole muualla Töölössä. Myös Kallion tarkastelussa lähekkäisten aiheiden, Helsingin ja Kallion, aihekennot olivat selkeästi päällekkäisiä laajalla alueella.

Töölön englanninkielisissä aiheissa fyysisessä kaupunkitilassa sijaitsevat kohteet, kuten Temppeliaukion kirkko ja Helsinki, näkyvät vahvemmin, kun taas suomenkielisissä aiheissa vain muutamassa aluekohtaisessa mallissa. Suomenlinna erosi aiheiltaan ja spatiaaliselta aiherakenteeltaan Kalliosta ja Töölöstä selkeästi, sillä kielien väliset aiheet olivat varsin samankaltaisia ja lähekkäin sijoittuneita. Tosin Suomenlinna on spatiaalisesti huomattavasti rajatumpi kaupunginosa verrattuna muihin, joten spatiaalinen lähekkäisyys ei siinä mielessä ole kovin yllättävää. Suomenlinnan aiheiden samankaltaisuus kielten välillä johtuu pitkälti siitä, että Suomenlinna on molemmille kieliryhmille pitkälti samankaltainen paikka: nähtävyys tai matkakohde. Tosin suomenkielisissä aiheissa näkyi myös paikallisuuteen liittyviä intressejä kuten häät ja eräs Suomenlinnassa asuva blogi-kirjoittaja, mutta myös englanninkielisten aiheiden kanssa yhteneväisyys oli ilmeistä. Suomenlinnan kieltenväliset samankaltaiset aiheet keskittyivät paikan, meriluonnon ja

maisemien ihannointiin. Kallion osalta kieltenväliset aihe-erot sijoittuvat Töölön selkeän erilaisuuden ja Suomenlinnan samankaltaisuuden väliin, tosin lähemmäksi Suomenlinnan samankaltaisuutta. Erityisenä huomiona Kallion aihemalleissa on arkisuuden näkyminen myös englanninkielisessä kaikki sanat huomioivassa aihemallissa työ-aiheen noustua näkyviin. Kenties taustalla ovat englanniksi julkaisevat nuoret suomalaiset (Leppänen et al. 2011), mutta taustalla voi olla Kallion alueella työskentelevät englanniksi julkaisevat muunmaalaiset Instagram-käyttäjät. Kalliossa yleiset ruokailuun ja sosiaaliseen kanssakäymiseen liittyvät aiheet ovat molemmilla kielillä kaikissa aihemalleissa näkyvissä vaikkakin hieman eri alueilla. Englanninkielisten aiheiden osalta Kalliossa lisäksi juhlitaan, jota ei näy suomenkielisissä aiheissa lainkaan. Suomenkielisissä aiheissa erona englanninkielisiin nautitaan auringonpaisteesta ja tehdään julkaisuja Kalliosta alueena ylipäätään.

Kielten välisten aiheiden yleispiirteisen samankaltaisuuden lisäksi aiheista löytyy eroja niitä tarkastellessa eri näkökulmista. Eroja on nähtävissä niin aiheista itsestään ja julkaisumäärien spatiaalisista eroista, mutta myös aiheiden ajallisesta ja spatiaalisesta rakenteesta. Englanninkielisen aineiston aiherakenne on ajallisesti tarkasteltuna suomenkielisestä aineistosta varsin selkeästi erottuva, mutta spatiaalisesti tarkasteltuna selkeät erot häivettyvät hieman. Koko Helsingin mittakaava ja siihen mallinnetut 10 aihetta tekivät spatiaalisesta tarkastelusta hankalan, mutta paljastivat yleispiirteistenkin aiheiden hajautuneen rakenteen. Pienemmät aluetarkastelut tarjosivat paremman tarttumispinnan aiheiden spatiaalisiin eroihin. Samojen tai samankaltaisten aiheiden päällekkäisyyttä on jonkin verran, erityisesti Suomenlinnan alueella, josta voidaan päätellä, että kielten välinen spatio-temporaalinen aiherakenne yhtenäistyy, kun alue on ryhmille samankaltainen, vaikka onkin englantivoittoinen. Käytetystä kielestä riippumatta, Suomenlinna on nähtävyys ja matkakohde, jolloin sen aiherakenne muillakin kielillä voisi olla varsin samankaltainen tässä työssä nähtyjen aiherakenteiden kanssa. Kallio on suomenkielisten julkaisujen hallinnoima alue, jossa suomeksi tehdään julkaisuja enemmän laajemmalta alueelta kuin englantia. Tämän lisäksi suomenkieliset aiheet ovat vahvemmin arkielämään kytköksissä kuin englanninkieliset, eikä alueella muodostu kolmipiikkistä julkaisurytmiikkaa. Töölön osalta se on englanninkielisessä aiherakenteessa nähtävyyksien ja Helsingin ihannoinnin alue, kielten väliset yhtymäkohdat liittyvät urheiluun ja tapahtumiin. Suomenkielinen

aiherakenne Töölössä on myös vahvemmin arkeen kytkeytynyt. Fyysinen kaupunkitila vaikuttaa ohjaavan julkaisujen aiheita kaupunginosatasolla vahvemmin kuin koko kaupungin tasolla, jota on havaittu myös muualla (Martin & Schuurman 2017).

4.3 Helsingin digitaalinen kaupunkitila näyttäytyy eri lailla kyseisten kieliryhmien julkaisuissa

Tämä tutkielma on osoittanut, että Helsingin digitaalinen kaupunkitila näyttäytyy eri lailla eri kieliä käyttävien kieliryhmien julkaisuissa ajallisesti, että alueellisesti. Teoriakirjallisuudessa käyty keskustelu digitaalisesta kaupunkitilasta (Kitchin 1998; Cohen 2007; Batty 2010; Graham & Zook 2011, 2013; Tranos & Nijkamp 2013; Kellerman 2014, 2015: Malecki 2017; Rose 2017; Zook & Graham 2017) sai empiiristä näyttöä tutkielman tuloksista. Ristiriitaiseksi, monimutkaiseksi ja jopa paradoksaaliseksi kuvailtu digitaalinen kaupunkitila (Graham & Zook 2011, 2013; Kellerman 2014; Malecki 2017; Rose 2017) näyttäytyi aihemallinnetuissa suomen- ja englanninkielisissä Instagram-julkaisuissa varsin selkeästi, vaikka tarkastelussa keskityttiinä pelkästään Instagram-julkaisujen kuvateksteihin. Instagram-julkaisujen vahvahko kytkös vapaa-aikaan molempien kielten osalta ei tarkoita, ettei vapaa-aikaan liittyvässä digitaalisessa kaupunkitilassa olisi sisällöllistä ristiriitaisuutta, kuten anakronistisia sanoja, vastakkaisia aiheita ja samankaltaisten aiheiden eriävää spatio-temporaalista rakennetta. Monikielisyyden huomioiminen entisestään monimutkaistaa digitaalisen kaupunkitilan luonnetta vaikeaselkoisena käsitteenä, mutta toimii samalla luontevana keinona jäsentää sitä.

Instagramin kautta välittyvä Helsingin digitaalinen kaupunkitila englanniksi julkaisevalle ryhmälle ei ole arjen ja viikonlopun rytmin hallitsema yhtä voimakkaasti kuin suomeksi julkaisevan digitaalinen kaupunkitila. Vuorokaudenajallinen julkaisuaktiivisuus taas paljasti digitaalisen kaupunkitilan rytmiikan olevan samanlainen kieliryhmien välillä yleisellä tasolla, mutta olevan monimuotoisempi paikallisella tasolla. Koko kaupungin tasolla vuorokaudessa on molemmilla kielillä kolme julkaisumääräpiikkiä samoihin aikoihin, mutta kaupunginosatasolle siirryttäessä selkeä kolmipiikkinen rakenne, Töölöä lukuun ottamatta, hälvenee

kaksi- ja yksipiikkiseksi, jolloin julkaisupiikit loivenevat ja painottuvat iltaan. Suomenlinnan julkaisuaktiivisuus on käytännössä piikitön.

Koko Helsingin mittakaavassa vertailtuna kieliryhmien välillä on samankaltaista julkaisumäärään perustuvaa spatiaalista rakennetta. Kieleen perustuvia spatiaalisia rakenteita on todettu myös muualla esimerkiksi Ranskassa, Kanadassa ja Israelissa (Graham & Zook 2013). Tiiviisti rakennetut asuin- ja palvelualueet keräävät suuren määrän julkaisuja molemmilla kielillä, vaikka maankäyttöä tarkastellessa kielten välillä oli nähtävissä jakautumista paikallisiin ja turisteihin. Englanninkielisiä julkaisuja oli esimerkiksi selkeästi enemmän lentokenttä- ja satama-alueilla kuin suomenkielisiä julkaisuja. Kielten välinen dominanssitarkastelu jakoi Helsingin keskustan kahtia suomen- ja englanninkielisten julkaisujen alueiksi Töölönlahden ja Kaisaniemen sillan kohdilta, englanninkielisten julkaisujen hallitessa eteläistä puoliskoa ja suomenkielisten pohjoista. Keskustan ulkopuolella suomenkieliset julkaisut olivat enemmistöä miltei kaikkialla.

Aiheiden kautta välittyvä kuva koko Helsingin digitaalisesta kaupunkitilasta on varsin monimuotoinen, mutta sen voi yksinkertaistaa havainnollisuuden vuoksi. Helsingin digitaalinen kaupunkitila näyttäytyy suomenkielisissä julkaisuissa pääasiallisesti arjen, sosiaalisen kanssakäymisen, juhlapyhien sekä kesän kautta. Englanninkielisissä julkaisuissa Helsingin digitaalinen kaupunkitila näyttäytyy sosiaalisen kanssakäymisen, juhlapyhien, Helsingin ihannoinnin ja eri aikojen, kuten aamun, talven ja viikonlopun, kautta. Koko Helsingin mittakaavan aihetarkastelun kautta välittynyt kuva digitaalisesta kaupunkitilasta kielten välillä on kahtiajakautunut, yleispiirteinen, monimuotoinen sekä hieman ristiriitainen. Yleispiirteisyyden vuoksi on erittäin todennäköistä, että Helsingin digitaalisen kaupunkitilan aiherakenne on todellisuudessa tässä esiteltyä monipuolisempi.

Kaupunginosatasolla ristiriitainen monimuotoisuus selkeytyy, sillä suurta alaa vastaavan digitaalisen kaupunkitilan sijasta aiheista välittyy vain kyseisen kaupunginosan digitaalinen kaupunkitila. Kaupunginosatason alueellisessa tarkastelussa aiheet tarkentuivat ja sitä myötä myös digitaalisen kaupunkitilan näyttäytyminen kielten välillä. Kaupunginosan “luonteesta” riippuen digitaalinen kaupunkitila

näyttäytyi joko eri lailla tai samankaltaisesti tarkastelussa olevilla kieliryhmillä. Helsingistä ei paljastunut yhtä vahvaa alueiden profiloitumista kuin mitä muualla on paljastunut (Lansley & Longley 2016; Martin & Schuurman 2017). Samankaltaisin alue kielten välillä tässä työssä tarkastelluista on Suomenlinna, joka ei ole kovin yllättävää, sillä se on molempien kieliryhmien edustajille samankaltainen alue: vapaa-ajan toimintoihin keskittyvä retkikohde. Digitaalisen kaupunkitilan Suomenlinna näyttäytyy molemmilla kielillä pääasiallisesti hienojen maisemien ja kesäpäivien retkikohteena, jolloin sitä tarkastellaan arjen ulkopuolisena eksoottisena paikkana (Urry & Larsen 2011, 13-18). Lisäksi julkaisumäärät ovat alueellisesti erittäin samankaltaisia kielten välillä vahvistaen digitaalisen kaupunkitilan näyttäytymistä samanlaisena kielestä riippumatta. Tosin suomenkielisten aiheidenn tarkastelussa Suomenlinnan vahva kytkös häiden järjestämispaikkana näkyi, kun taas englanninkielisessä tarkasteluissa se ei näkynyt lainkaan.

Töölön osalta kieliryhmien välisiä eroja oli Suomenlinnaa selkeästi enemmän digitaalisessa kaupunkitilassa, vaikkakin pientä samankaltaisuuttakin on havaittavissa. julkaisumääriä tarkastellessa digitaalisen kaupunkitilan rakenne oli samankaltainen kielten välillä, tosin englanninkielisessä aineistossa tietyt nähtävyydet korostuivat selkeästi vahvemmin kuin suomenkielisessä aineistossa. Töölössä on molemmilla kielillä urheilu- ja tapahtuma-aiheiden osalta samankaltainen digitaalinen kaupunkitila. Suomenkielisten aiheiden kautta Töölön digitaalinen kaupunkitila eroaa englanninkielisestä vuodenaikojen vaihtelun, valokuvauksellisuuden ja sosiaalisen kanssakäynnin kautta, kun taas englanninkielisten aiheiden kautta Töölö näyttäytyy enemmän ruokailun, nähtävyyksien, Suomen ja Helsingin ihailun kautta.

Kallion digitaalinen kaupunkitila vaikuttaa julkaisumäärän perusteella olevan erilainen kielten välillä, sillä painopisteet vaikuttavat olevan osittain eri paikoissa. Toisaalta aiheiden perusteella Kallio vaikuttaa olevan Suomenlinnan ja Töölön välimaastossa sen suhteen, mitä tulee digitaaliseen kaupunkitilaan. Molemmilla kielillä Kallio näyttäytyy kahvittelun, työn, ruokailun ja vapaa-ajan alueena, mutta englanninkielisen aineiston kautta Kallio näyttäytyy myös juhlimiseen liittyvänä alueena, joka puolestaan ei näy suomenkielisessä aineistossa. Toisin sanoen, Kallio näyttäytyy aiheiden osalta melko samankaltaisena, mutta

julkaisumäärällisesti erilaisena eli samankaltaisista aiheista riippumatta kieliryhmien digitaalinen kaupunkitila vaikuttaa eroavan siten, että saman aiheeseen liittyvät julkaisut tapahtuvat eri paikoista.

Mitä tulee aiheiden tarkasteluun Jan Gehlin kehikon (taulukko 1) kautta, tuloksien valossa on selvää, että lähes kaikki Instagram-julkaisut liittyvät valinnaisiin ja sosiaalisiin aktiviteetteihin, joten Instagramin kautta näyttäytyvä digitaalinen kaupunkitila tästä näkökulmasta tarkasteltuna on myös vahvasti kytköksissä vapaa-aikaan, eikä siinä siten näy juurikaan välttämättömiä aktiviteetteja. Ympäristölaatukehikko ei siis tuo Instagram-aineistojen kautta näkyvän digitaalisen kaupunkitilan analysointiin kovinkaan käyttökelpoista lisäulottuvuutta, mutta se voisi toimia muiden sosiaalisen median alustojen aineistoilla tai usealta alustalta yhdistetyllä aineistolla (ks. Citizen Mindscapes 2018).

Digitaalisen kaupunkitilan näyttäytyminen englanniksi ja suomeksi Instagram-julkaisuja tekevien välillä on siis kaiken kaikkiaan osittain samankaltaista, mutta selkeitä eroja löytyy alueellisesti aiheita ja julkaisumääriä vertaillessa. Fyysinen kaupunkitila heijastuu digitaaliseen kaupunkitilaan julkaisujen aiheiden melko hyvin huolimatta aiheiden monimuotoisesta ja ristiriitaisestakin spatio-temporaalisesta rakenteesta. On lähes itsestään selvää, että Instagram-julkaisuaineiston sisällön tuottaneen suuren käyttäjäjoukon mielipiteiden, asenteiden ja näkökulmien kirjo on laaja, mutta on silti yllättävää, että tämä kirjo näyttäytyi tuloksissa pelkästään julkaisujen kuvateksteihin keskittyneen analyysin kautta. Oletettavissa on, että kuvien, kuvasarjojen, videoiden ja kommenttien sisällytys analyysin olisi tuonut tämän kirjon vielä kirkkaammin esille.

Tämänkaltainen spatiaaliseen sosiaaliseen mediaan keskittyvä kaupungin tarkastelutapa konkretisoi siellä asuvien ja liikkuvien ihmisjoukkojen monimuotoisuutta ja siihen monimuotoisuuteen olennaisena osana kuuluvaa ristiriitaisuutta. Kieliryhmien välisiä eroja digitaalisen kaupunkitilan näyttäytymisessä voisi verrata vastaavalaiseen fyysisen kaupunkitilan tarkasteluun esimerkiksi nuoren ja vanhan ihmisen välillä: heitä kiinnostavat todennäköisesti eri asiat ja paikat fyysisessä kaupunkitilassa, mutta yhteneväisyyksiäkin löytynee. Tämän työn tuloksissa ja työn eri vaiheissa ilmenneet havainnot digitaalisen kaupunkitilan ominaisuuksista konkretisoivat teoreettista keskustelua digitaalisesta kaupunkitilasta ja kybermaisemista,

joka on erilainen paikka eri ihmisille ja jota eri ihmiset muokkaavat näköisekseen (Kellerman et al. 2015; Malecki et al. 2017; Rose 2017). Tosin tulee muistaa, että Instagram on vain yksi sosiaalisen median alusta lukuisten alustojen joukossa ja, että työssä ei keskitytty muuhun kuin kuvatekstien kautta välittyvään kaupunkitilaan. Katsausta kaupunkitilaan voisi laajentaa huomattavasti huomioimalla myös julkaisujen kuva- ja videosisällön, sekä mahdollisesti myös julkaisujen saamat muiden käyttäjien kirjoittamat kommentit. Silloinkin kyse olisi vain Instagramin kautta näkyvästä digitaalisesta kaupunkitilasta. Kokonaisvaltainen digitaalisen kaupunkitilan tutkiminen ja ymmärtäminen vaatisi usean sosiaalisen median alustan julkaisujen tarkastelua monipuolisella menetelmävalikoimalla, jotka yhdistelisivät paikkatietoa, kieliteknologisia menetelmiä sekä konenäköä. Näiden lisäksi, tässä työssä tarkasteltu digitaalinen kaupunkitila on sellainen, joka ei välttämättä näy Instagram-käyttäjille itselleen, vaikka se on heidän itsensä tuottama. Jotta pääsisi siihen käsiksi, miten digitaalinen kaupunkitila näyttäytyy siellä liikkuvalle tulisi tarkastella esimerkiksi miten käytetty kieli tai käyttäjäprofiilin asetukset vaikuttavat hakukoneiden tarjoamiin näkymiin läheisistä ravintoloista tai reittivalinnoista.

4.4 Kieliteknologian menetelmät soveltuvat kaupunkitilan tutkimukseen

Kieliteknologian hyödyntäminen maantieteellisessä tutkimuksessa on melko uusi tapa lähestyä maantieteellisiä kysymyksiä, joka yleistyy jatkuvasti (Graham & Zook 2013; Jenkins et al. 2016; Lansley & Longley 2016; Martin & Schuurman 2017; Fu et al. 2018; Hiippala et al. 2018), mutta on Suomessa vielä varsin harvinaista. Tässä työssä pääasiallisena menetelmänä käytetty paikkatietoon yhdistetty LDA-aihemallinnus vaikuttaa soveltuvan melko hyvin myös maantieteellisen tutkimuksen välineeksi. Tekstiaineiston analysoiminen laskennallisesti ja tilastollisesti vaatii melko paljon esikäsittelyä, jotta analyysien tuloksena saatava tieto olisi käyttökelpoista ja täysin esikäsittelemättömän tekstin analysointi vastaa pitkälti “roskaa sisään, roskaa ulos” -periaatetta. Monikielisellä aineistolla kielentunnistuksen luotettavuus täytyy olla tarpeeksi korkea ja tunnistettavan tekstin tarpeeksi pitkä, jotta tunnistus on luotettava. Myös tulosten jälkikäsittely on tärkeää, kuten tässä työssä aiheisiin luokiteltujen julkaisujen suodattaminen, sillä selkeästi epävarmojen tuloksien rajaaminen pois vahvistaa lopputuloksen luotettavuutta. Toisin sanoen, on tärkeää säilyttää kriittinen suhtautuminen aineistoa kohtaan sekä

suodattaa aineiston epävarmuutta pois ennen analyysiä ja sen jälkeen, jotta tuloksista pystyy johtamaan oikeita johtopäätöksiä.

Työn tulokset osoittavat aihemallintamisen tuottavan parempia tuloksia silloin, kun mallille syötettävät dokumentit, joiden perusteella mallintaminen tehdään, ovat rajattu esimerkiksi maantieteellisesti. Lisäksi, aihekennot osoittautuivat hyväksi tavaksi käsitellä ja visualisoida aiheita spatiaalisesti. Oletettavasti myös ajallinen rajaus parantaisi aiheiden koherenssia ja siten myös tulkitsemisen helppoutta. Instagram-kuvatekstien osalta myös adjektiivien jättäminen mallintamisen ulkopuolelle vaikutti terävöittävän aihemallin tuloksia, jolloin aiheiden nimeäminen helpottui hieman. Suurelle tekstiaineistolle ilman rajauksia ja pienellä aihemäärällä tehty aihemallinnus tuottaa erittäin yleispiirteisiä aihekokonaisuuksia, jotka voivat olla semanttisesti lähekkäin toisiaan ja erityisen hankalia nimetä. Mitään selkeätä parasta toimintatapaa aihemäärien ja rajauksien suhteen ei ole vielä muodostunut, mutta tämän työn perusteella voi sanoa pienemmän aihemäärän riittävän, mikäli aineisto on rajattu pienehkölle alueelle tai aikavälille. Aineistolle sopivan aihemäärän löytyminen vaatii aihemallin ajamista useaan otteeseen eri parametreilla, kunnes sopivat parametrit löytyvät. Mikäli tutkijalla on aineistosta ennakkotietoja, niitä tulisi hyödyntää aihemallin parametreja määriteltäessä.

Monikielisyys asettaa selkeitä haasteita kieliteknologian menetelmien soveltamiseen. Suomenkielisen tekstin käsittely aihemallinnuksen kaltaisilla menetelmillä vaatii mittavan esikäsittelyn, jotta edes yksinkertaiset sanapilvet kuvaisivat aineiston sanastoa riittävän hyvin. Rikkaan morfologiansa vuoksi suomenkielisen tekstin yhdenmukaistaminen lemmatisoinnilla on välttämätöntä, kun taas englanninkieliselle tekstille se ei ole aivan yhtä välttämätöntä, mutta suositeltavaa. Suomi on lisäksi varsin pieni kieli, joten kielelle kehitettyjen menetelmien ja avoimien työkalujen valikoima on varsin suppea. Tässä työssä suomenkielisen tekstin lemmatisointiin käytetty FinnPOS-työkalu vaikeutti työtä, sillä osa työvaiheista tuli suorittaa Linux-käyttöjärjestelmässä työkalun vaatimusten vuoksi. Avoimille ja helppokäyttöisille suomenkieliselle tekstille soveltuville kieliteknologisille menetelmille ja työkaluille on tämän työn

perusteella tarvetta. Myös puhekielisten sanojen lemmatisointiin soveltuvaa mallia tarvitaan, sillä nyt puhekieliset ilmaisut ja sanat aiheuttivat pieniä ongelmia aihemallinnuksen tuloksien analysoinnissa.

Jatkoa ajatellen LDA-aihemallinnuksen soveltaminen aluekohtaisesti tai/ja ajanjaksokohtaisesti voi tarjota erittäin hyvän menetelmän tarkastella spatio-temporaalisia aiheita. Tiettyyn kaupunginosaan keskittyvä ja liikkuvaa aikaikkunaa hyödyntävä aihemallinnus paljastaisi aiheiden spatio-temporaalisen dynaamisuuden. Tällöin digitaalisessa kaupunkitilassa tapahtuvaa keskustelua ja tapahtumia pystyisi seuraamaan oletettavasti varsin hyvin. LDA-aihemallin esiparametreja voisi tämän työn tuloksien perusteella muuttaa oletusarvoista, kun sitä soveltaa Instagram-aineistolle, sillä alhaisen koherenssipisteiden ja yleispiirteisten aiheiden kautta voisi olettaa aiheiden jakautuvan epäsymmetrisesti keskenään. Myös muita aihemallintamismenetelmiä voisi soveltaa maantieteelliseen sosiaalisen median tutkimukseen kuten hierarkkista aihemallinnusta (engl. *hierarchical topic modeling*, Blei et al. 2003b), lyhyisiin teksteihin erikoistunutta aihemallinnusta (engl. *short text topic modeling*, Ramage et al. 2010; Mehrotra et al. 2013; Mazarura et al. 2014) ja LDA-aihemallin muokatut versiot kuten kirjoittajakohtainen (engl. *author-topic model*, Rosen-Zvi et al. 2004) ja ohjattu LDA (engl. *semi-supervised*, Ramage et al. 2010; Mehrotra et al. 2013). Tavallinen “muokkaamaton” LDA-aihemallinnus mallintaa sille syötetyistä teksteistä etenkin pienillä aihemäärillä varsin yleispiirteisiä aiheita, sillä se pyrkii sovittamaan kaikki tekstit jonkin aiheen alle. Tavallisella LDA-mallilla massasta poikkeavien aiheiden mallintaminen vaatii joko mallinnettavien aihemäärien nostamista aineistosta riippuen useisiin kymmeniin tai satoihin aiheisiin. Tarkempien aihekokonaisuuksien löytämiseen LDA:n lisäksi voisi kokeilla eri vektorirepresentaation hyväksyvää aihemallintamismenetelmää. LDA-aihemalli toimii oikein vain sanojen bag-of-words -vektorirepresentaatioiden kanssa, joissa jokaisella sanalla on oma tunnusluku sekä sanan esiintymistiheyttä kuvaava kokonaisluku, joka on varsin yksinkertainen vektorirepresentaatio. Esimerkiksi word2vec- tai fastText -vektorirepresentaatiot voisivat toimia varsin hyvin, sillä niiden mahdollistama sanojen samankaltaisuuden arviointi voisi tuoda aihemallinnukseen lisää ketteryyttä ja tarkkuutta, sekä mahdollistaisi neuroverkkopohjaisen aihemallintamisen (Larochelle & Lauly 2012; Esposito et al. 2016; Arslan et al. 2018).

Eri kielille mallinnettujen aiheiden välisten erojen tilastollinen kvantifioiminen on varsin hankalaa menetelmällisesti, sillä se vaatii samaa asiaa tarkoittavien eri kielisten sanojen vektorirepresentaatioiden olevan lähellä toisiaan. Se vaatisi vektorisointimenetelmän kehittämistä, joka olisi tietoinen sanojen merkityksistä ja siten osaisi määritellä, että sanojen ‘järvi’ ja ‘lake’ vektorit muistuttaisivat matemaattisesti toisiaan, kun taas sanojen ‘järvi’ ja ‘brewery’ olisivat matemaattisesti selkeästi erilaisia. Tällainen tarkastelutapa mahdollistaisi tehokkaan ja kvantitatiivisen aiheiden samankaltaisuuden arvioinnin eri kielten välillä. Vaihtoehtoisesti, tämän voisi toteuttaa ennen vektorointia tapahtuvalla sanojen automaattisella kääntämisellä yhteiselle kielelle, mutta se vaatisi tarkkaa kontekstiriippuvaista kääntämistä, jotta sanonnat, sekä monitulkintaiset sanat kääntyisivät oikein. Kyseisten menetelmien kehittäminen on tämän tutkielman kirjoittajan kirjoitushetken näkemyksen mukaan toisen opinnäytetyön arvoinen työ esimerkiksi lingvistiikassa tai tietojenkäsittelytieteessä, mutta esimerkiksi Googlen kehittämällä BERT-mallilla mainitun kaltainen toiminnallisuus on lähempänä käyttöönottoa (Devlin & Chang 2018). Kielitieteellisten teknologioiden laaja-alaisen omaksumisen ja kehittämisen osalta eräs ongelma-alue on menetelmäkehittämisen ja tutkimuksen pirstaleisuus eri tieteenalojen alle, jolloin parhaaksi osoittautuneet käytännöt ja menetelmät eivät ole välttämättä tutkijan tiedossa (Jauhiainen et al. 2018).

Työssä kokeiltu LDA-aihemallinnuksen soveltaminen kevyeen “sävyanalyysiin” adjektiiveilla ei toiminut lainkaan tällä aineistolla. Jotta aihemallinnus tuottaisi edes jollain tasolla käyttökelpoisia “sävyanalyysin” tuloksia, se vaatisi monipuolisemman sävysisällön aineiston. Tässä työssä muodostettujen adjektiivi-aihemallien aiheiden semanttinen erotteleminen toisistaan merkityksellisellä tasolla osoittautui mahdottomaksi adjektiivien positiivisuuden vuoksi ja malleista oli nähtävissä, että tulokset eivät toisi analyysiin toivottua lisäsyvyyttä vaan täysin irrallisen ja käyttökelvottoman muuttujan, joten Occamin partaveitsen periaatteen mukaisesti ne hylättiin. Sävyanalyysi tulee toteuttaa varta vasten siihen kehitetyillä työkaluilla eikä esimerkiksi aihemallinnusmenetelmillä, jotta tuloksista voisi vetää johtopäätöksiä ja tuloksia ylipäätänsä voisi kvantifioida numeerisesti. Sävyanalyysin soveltaminen Instagram-aineistoon vaikuttaa tämän työn löydöksien kautta mielenkiintoiselta, sillä aineisto vaikuttaa olevan erittäin vahvasti positiivissävytteisiä ja negatiivisten julkaisujen löytyminen olisi jo itsessään kiinnostavaa.

4.5 Lopuksi

Sosiaalisessa mediassa tuotetun massadatan monipuolisuus tekee siitä erityisen rikkaan, mutta samalla hankalan ja oikukkaan, aineistolähteen. Ensinnä, tässä työssä on keskitytty vain Instagram-alustalla tehtyjen julkaisujen kuvateksteihin Helsingistä, eikä esimerkiksi kuviin tai kommentteihin, mutta niinkin rajatussa tarkastelussa sosiaalisen massadatan monimuotoinen ja ristiriitainen luonne paljastui. Kuvatekstit voivat sisältää lähes mitä vain: lyhyitä ja pitkiä kuvatekstejä, useita kieliä, aihetunnisteita, mainintoja, emojeita, hymiöitä ja hyperlinkkejä, sekä yliampuvaa välimerkkien käyttöä. Tämän vuoksi sosiaalisen median tekstuaalisia aineistoja täytyy esikäsitellä ennen analysointia, jolloin osa rikkaudesta auttamatta häviää.

Toiseksi, esikäsittelemisessä on tiedettävä mitä tietoa tekstistä haluaa saada, jolloin tiettyjen elementtien sisällytys on oleellista, esimerkiksi julkaisujen tekstien sävyjen kannalta hymiöt ja emojit ovat tärkeää tietoa. Esikäsitteleminen on kaiken lisäksi aikaa vievä työnvaihe ja tämän työn työtunneista erittäin suuri osa kului pelkästään aineiston esikäsittelemisessä aihemallinnukselle sopivaan muotoon.

Kolmanneksi, kohdepisteisiin perustuvat sijainnit tuovat mukanaan epävarmuustekijän geoleimauksien tarkkuuden kanssa. Julkaisujen pakollinen geoleimaaminen jonkin kohdepisteen alle tekee aineistoon taipumuksen suosia tiettyjä kohdepisteitä toisten sijaan, jolloin tietyt pisteet haalivat suuren määrän julkaisuja itselleen, jolloin julkaisujen todellinen melko hienojakoinen spatiaalinen rakenne ei heijastu aineistoon samalla spatiaalisella tarkkuudella. Tämän aggregaation vaikutus korostuu pinta-alaltaan rajattuja alueita tarkastellessa, kuten Suomenlinnan alueella. Lisäksi, tiettyyn sijaintiin liitetyt julkaisut eivät välttämättä käsittele kohdepisteen sijaintia lainkaan, eikä käyttäjä julkaisuaan tehdessä välttämättä näe kohdepisteestä muuta tietoa kuin nimen, eikä esimerkiksi nimeä ja sijaintia kartalla, jolloin käyttäjällekin on epävarmaa mihin sijaintiin julkaisu oikeastaan sidotaan. Esimerkiksi julkaisu, joka on sidottu Helsingin musiikkitalon kohdepisteeseen ei välttämättä käsittele taloa tai talon sisällä tapahtuvia asioita, vaan piste saattoi vain olla käyttäjää lähinnä oleva piste hänen tehdessä julkaisua jostain täysin muusta asiasta.

Neljänneksi, aineiston oikukkuus näkyy lisäksi Instagram-julkaisujen kuvatekstien aiheiden “ohuen” monipuolisuuden kautta, sillä yleispiirteisten aiheiden sisälle mahtuu useita julkaisuja, mutta aiheiden

tarkentuessa niiden alle luokittuvien julkaisujen määrä pienenee. Näin ollen, sosiaalisen median aineistoja käytettäessä tutkimusaineistona tulee valita tarkkaan tutkimukseen soveltuvat sosiaalisen median alustat ja huomioida, että tarkasti määriteltyihin kysymyksiin vastauksia tarjoavia julkaisuja ei välttämättä ole millään alustalla kovinkaan paljon, etenkin kun julkaisuja lähtee suodattamaan esimerkiksi kielen ja tekstin pituuden kautta. Tämän tutkimukseen sopivien julkaisujen määrän niukkuuden lisäksi, sosiaalisen median aineistojen kanssa on erittäin vaikea arvioida etukäteen, pystyykö kyseinen aineisto antamaan vastauksia haluttuihin tutkimuskysymyksiin ennen kuin aineiston on saanut esikäsiteltyä ja aineistosta on saanut alustavia tuloksia. Tämän vuoksi vaikuttaakin, että suuri osa sosiaalista mediaa hyödyntävästä tutkimuksesta on luonteeltaan eksploratiivista ja usein sen ohessa toteutetaankin jokin tavanomaisempi analyysi, jota sosiaalisen median analyysin tulokset syventävät.

Viidenneksi, käytäntöä läheisempi ja korkeamman tason oikukkuus on erittäin todennäköinen mahdollisuus, että sosiaalisen median rajapintapalvelut ja käyttöehdot muuttuvat ajan myötä, jonka jälkeen rajapinnalle kehitetyt työkalut voivat lakata toimimasta, rajapintapalvelut voidaan poistaa ja aineiston kerääminen käyttämällä jotain muuta menetelmää kuin rajapintapalveluita on usein käyttöehtojen vastaista, joiden rikkominen voi johtaa oikeudenkäyntiin (Freelon 2018).

Näistä haasteista huolimatta sosiaalisen median aineistot ja muut massadata-aineistot antavat erinomaisen mahdollisuuden tarkastella ja ymmärtää nykyaikaista maailmaa. Nämä aineistot ovat todennäköisesti alati keskeisempää roolia modernien yhteiskuntien, kulttuurien ja ihmisten tutkimisessa ja ymmärtämisessä, sillä sosiaalisen median käyttäjät ja massadatan määrät jatkavat kasvuaan (Statista 2018b, 2018d). Aineistoista saatua ymmärrystä ja tietoa voidaan hyödyntää esimerkiksi selvitettäessä minkälaisia aktiviteetteja ja puheenaiheita tietyt alueet keräävät tietyn ihmisryhmän osalta. Jotta sosiaalisen median aineistojen tuloksia voisi yleistää koskemaan kattavampaa joukkoa yhteiskunnasta, aineistot tulisi kerätä usealta alustalta, sillä eri sosiaalisen median alustojen käyttäjäkunnat ovat erilaisia. Joka tapauksessa, sosiaalisen median aineistot vaikuttavat tarjoavan arvokkaan aineiston alueiden, kaupunkien, yhteiskuntien ja kulttuurien maantieteelliseen tutkimukseen sekä soveltuvan oheisaineistoksi päätöksenteon tukemiseen.

Kieliteknologian menetelmät ja muut koneoppimiseen liittyvät menetelmät voivat tuoda maantieteelliseen tutkimukseen erittäin arvokkaan lisän. Laajojen kvalitatiivisten tekstiaineistojen nopea ja tehokas sisällönanalyysi tekstissä olevien sävyjen, piilevien aiheiden, paikkaviitteiden ja muiden piirteiden löytämiseksi mahdollistaa maantieteellisten kvalitatiivisten aineistojen kvantifioimisen uusin keinoin. Täten syntyvä tieto, ymmärrys ja tutkimussuuntaukset voivat viedä maantieteellistä tutkimusta ja tieteellistä ymmärrystä maailmastamme uusiin suuntiin (Martin & Schuurman 2017), mutta se vaatii vielä melko paljon työtä menetelmällisten ja käytännön ongelmien selvittämiseksi.

Suomenkielisen tekstin osalta kieliteknologiset menetelmät, kuten lemmatisointi, ovat selkeästi hankalampia toteuttaa verrattuna englantiin. Suomi on kielenä morfologisesti erittäin rikas kieli, jolloin sanoilla on lukuisia eri taipuneita muotoja, joka tekee suomen käsittelyyn sopivien menetelmien kehittämisestä erityisen hankalaa. Lisäksi suomi on maailman mittakaavassa varsin pieni kieli, jonka vuoksi kiinnostus ja kysyntä tämänkaltaisia menetelmiä kohtaan on vähäistä verrattuna esimerkiksi englantiin eikä suomelle kehitettyjen työkalujen kirjo ole kovinkaan suuri. Suomenkielisten tekstiaineistojen automaattisen käsittelyn mahdollistaminen kieliteknologian menetelmillä on erityisen tärkeässä osassa tulevaisuutta ajatellen, mikäli suomenkielisiä aineistoja kohtaan on maantieteellisen tutkimuksen osalta pyrkimyksiä. Esimerkiksi sävyanalyysien vaatimien mallien luominen, kirja- ja puhekielen välisten erojen vaikutusten minimoiminen ja helppokäyttöisten avoimien työkalujen kehittäminen muodostaisivat erinomaiset edellytykset suomenkielisen tekstin käsittelylle lingvistisillä tekniikoilla. Kuten tässä työssä todettiin kieliteknologian menetelmät soveltuvat varsin hyvin maantieteelliseen tutkimukseen ja vaikuttavat mahdollistavan täysin uusia lähestymistapoja maantieteellisiin kysymyksiin. Sävyanalyysien lisäksi erittäin mielenkiintoinen menetelmä on geojäsentäminen (engl. *geoparsing*), jossa tekstissä sijaitsevat paikannimet tunnistetaan ja geokoodataan paikkatiedoksi. Geojäsentäminen mahdollistaa esimerkiksi paikannimien alla “piilevän” paikkatiedon louhimisen mistä tahansa tekstistä, mutta menetelmässä on epävarmuustekijöitä, jotka tulee huomioida (Gritta et al. 2018).

Työn tuloksia ja työssä käytettyjä menetelmiä voi hyödyntää muun muassa digitaalisen kaupunkitilaan liittyvissä jatkotutkimuksissa ja teoreettisessa keskustelussa, markkinoinnin kohdistamiseen oikeisiin

paikkoihin oikeilla kielillä ja kaupunkisuunnittelussa. Sosiaaliseen mediaan tallentuneiden aiheiden, sijaintien ja aikojen kautta pystyy muodostamaan näkymän digitaaliseen kaupunkitilaan ja siten myös oikeaan kaupunkitilaan. Geoleimauksessa tallentuvaa paikka- ja aikatietoa pystyy hyödyntämään selvitettäessä kaupungissa liikkuvien henkilöiden oikeita käyntikohteita ja -aikoja. Näillä tiedoilla voi saada arvokasta tietoa alueiden tai paikkojen kävijöistä etenkin, jos kävijämääriä ja -kokemuksia ei entuudestaan tiedetä. Sosiaalisen median datan parhaimpana ja huonoimpana puolena on sen “ohut monipuolisuus”, sillä siellä keskustellaan lukuisista asioista mutta samalla ei välttämättä mistään tietystä asiasta. Tämän vuoksi kokonaisvaltaisen ymmärryksen kannalta on olennaista tarkastella sosiaalisen median ilmiöitä, sekä digitaalista kaupunkitilaa, usealta alustalta käsin ja erilaisia menetelmiä käyttäen. Esimerkiksi vielä on epäselvää, mikä suhde Instagram-julkaisun kuvatekstillä on julkaisun kuvan aiheen kanssa ja tämä kysymys muodostaakin erityisen olennaisen kysymyksen mihin sosiaalisen median analyyseissä tulisi kiinnittää huomiota.

Tietokoneen laskentatehoa ja tekoälyä hyödyntävät lingvistiset menetelmät vaikuttavat sopivan hyvin maantieteelliseen tutkimukseen sosiaalisesta mediasta, mutta niiden yhdistäminen muihin tekoälyä hyödyntäviin sisällönanalyysin menetelmiin, kuten konenäköön, voisi vahvistaa käyttökelpoisuutta entisestään. Aihemallinnuksen kaltaiset kvantitatiivisia ja kvalitatiivisia piirteitä yhdistelevät menetelmät voivat saada aikaan melko suuren mullistuksen erityisesti kulttuuri- ja ihmismaantieteen tutkimuksen puolella mahdollistaen suurten aineistomäärien nopean käsittelyn ja analysoinnin. Paikkatietomenetelmien ja tekoälyn yhdistäminen mahdollistaa spatiaalisen mallintamisen viemisen aivan uudelle tasolle, jossa malli paranee uusista havainnoista oppiessaan. Tekoälyn avustamia menetelmiä ja poikkitieteellistä lähestymistä spatiaalisten ilmiöiden tutkimiseen tarvitaan jatkossa, erityisesti massadata-aineistojen kanssa. Digitaalinen kaupunkitila elää ja muuttuu, sekä siihen vahvasti kytköksissä olevat massadata-aineistot kasvavat määrällisesti niin kauan kuin internetiä ja oikean maailman sijainteja hyödyntäviä verkkosovelluksia ja sosiaalisen median alustoja käytetään. Jotta maantieteellinen tutkimus pysyisi nopeasti muuttuvan ja digitalisoituvan maailman mukana, tässä työssä käytettyjä menetelmiä ja uusia menetelmiä tulee soveltaa kasvavissa määrin kulttuurin, yhteiskunnan ja ympäristön, maailmamme, ymmärtämiseksi.

5.0 Kirjallisuus

Aletras, N., T. Baldwin, J. H. Lau & M. Stevenson (2014) Representing Topic Labels for Exploring Digital Libraries. *Proceedings of the 14th IEEE/ACM Joint Conference on Digital Libraries*, 239-248. *<DOI:10.1109/JCDL.2014.6970174>* Viitattu 17.10.2018.

Arslan, Y., D. Kucuk & A. Birturk (2018) Twitter Sentiment Analysis Experiments Using Word Embeddings on Datasets of Various Scales. *23rd International Conference on Applications of Natural Language to Information Systems*, 40-47. *<DOI:10.1007/978-3-319-91947-8_4>*

Axelbrooke, Stuart (2018) LDA Alpha and Beta Parameters *<https://www.thoughtvector.io/blog/lda-alpha-and-beta-parameters-the-intuition/>* Viitattu 05.05.2018

Batty, Michael (2010) The Pulse of the City. *Environment and Planning B: Planning and Design* 37, 575-577. *<DOI:10.1068/b3704ed>*

Bendler, J., S. Wagner, T. Brandt & D. Neumann (2014) Taming Uncertainty in Big Data, Evidence from Social Media in Urban Areas. *Business & Information Systems Engineering* 5, 279-288.

Birch, C. P. D., S. P. Oom & J. A. Beecham (2007) Rectangular and hexagonal grids used for observation, experiment, and simulation in ecology. *Ecological Modelling* 206: 3-4, 347–359.

Blei, D., A. Y. Ng & M. I. Jordan (2003a) Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 993-1022.

Blei, D. M., T. L. Griffiths, M. I. Jordan & J. B. Tenenbaum (2003b) Hierarchical Topic Models and the Nested Chinese Restaurant Process. *Teoksessa* Thrun, S., K. Saul & B. Scholkpof (toim.) *Advances in Neural Information Processing Systems 16*, 17-24.

Blei, David (2012a) Probabilistic Topic Models. *Communications of the ACM* 55: 4, 77-84.

Blei, David (2012b) Topic Modeling and Digital Humanities. *Journal of Digital Humanities*. *<http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/>* Viitattu 18.10.2018.

Brants, Keer (2006) Guest Editor's Introduction: The Internet and the Public Sphere. *Political Communications* 22: 2, 143-146. <DOI:10.1080/10584600590933133>

Cadwalladr, Carole & Emma Graham-Harrison (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *<https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>* Viitattu 10.08.2018.

Carter, S., W. Weerkamp & M. Tsagkias (2013) Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation* 47: 1, 195–215. *<DOI:10.1007/s10579-012-9195-y>*

Castree, N., R. Kitchin & A. Rogers (2013) Oxford Dictionary of Human Geography. 1. painos, 572 s. Oxford University Press, Oxford, Iso-Britannia.

Celebi, Arda & Arzucan Ozgur (2017) Segmenting hashtags and analyzing their grammatical structure. *Journal of The Association For Information Science and Technology* 69: 5, 675-686.

Chen, Edwin (2011) Introduction to Latent Dirichlet Allocation. *<http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/>* Viitattu 04.04.2017

Chung, C., E. Agapie, J. Schroeder, S. Mishra, J. Fogarty & S. A. Munson (2017) When Personal Tracking Becomes Social: Examining the Use of Instagram for Healthy Eating. *Proceedings of the 2017 ACM SIGHCI Conference on Human Factors in Computing Systems*, 1673-187.

Citizen Mindscapes (2018) Citizen Mindscapes -tutkijakollektiivi. *<https://blogs.helsinki.fi/citizenmindscapes/tutkijakollektiivi/>* Viitattu 19.11.2018.

Coary, Sean & Morgan Poor (2016) How consumer-generated images shape important consumption outcomes in the food domain. *Journal of Consumer Marketing* 33: 1, 1-8.

Cohen, Julie (2007) Cyberspace as/and Space. *Columbia Law Review* 107: 210, 210-256.

Crandall, D., L. Backstrom, D. Huttenlocher & J. Kleinberg (2009) Mapping the World's Photos. *Proceedings of 18th International Conference on World Wide Web*, 761-770.

Cvetojevic, S., L. Juhász & H. H. Hochmair (2016) Positional Accuracy of Twitter And Instagram Images in Urban Environments. *GI_Forum* 1, 191-203.

Deshpande, Devashish (2018) What is Topic Coherence? *<https://rare-technologies.com/what-is-topic-coherence/>* Viitattu 02.05.2018

Devlin, Jacob & Ming-Wei Chang (2018) Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. *<https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>* Viitattu 6.11.2018.

Dotson, Taylor (2012) Technology, choice and the good life: Questioning technological liberalism. *Technology in Society* 34, 326-336.

Drucker, Susan J. & Gary Gumpert (2012) The Impact of Digitalization on Social Interaction and Public Space. *Open House International* 37: 2, 92-99.

Eksymä, Anna (2009) Puolustusvoimien saarten muuttaminen siviilikäyttöön hidasta. *<https://yle.fi/uutiset/3-5255668>* Viitattu 18.11.2018.

Esposito, F., A. Corazza & F. Cutugno (2016) Topic Modelling with Word Embeddings. *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it*, 129-134. *<DOI:10.1109/ISCC.2017.8024509>*

ESRI (2018) Why Hexagons? *<http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-whyhexagons.htm>* Viitattu 20.01.2018

ExplosionAI (2018) spaCy: Industrial Strength Natural-Language Processing in Python. <*https://spacy.io*> Viitattu 14.11.2018

Freelon, Deen (2018) Computational Research in the Post-API Age. *Political Communication*, 1-4. <*DOI:10.1080/10584609.2018.1477506*>

Fu, C., G. McKenzie, V. Frias-Martinez & K. Stewart (2018) Identifying spatiotemporal urban activities through linguistic signatures. *Computers, Environment and Urban Systems* 72, 25-37. <*DOI:10.1016/j.compenvurbsys.2018.07.003*>

Gehl, Jan (2011) *Life Between Buildings: using public space*. 6. p. 207 s. Island Press, Washington D.C., Yhdysvallat.

Graham, Mark & Matthew Zook (2011) Visualizing Global Cyberscapes: Mapping User-Generated Placemarks. *Journal of Urban Technology* 18: 1, 115-132.

Graham, Mark & Matthew Zook (2013) Augmented realities and uneven geographies: exploring the geolinguistic contours of the web. *Environment and Planning A* 45: 1, 77–99.

Greenwood, S., A. Perrin & M. Duggan (2016) Social Media Update 2016. Pew Research Centre, 1-19.

Gritta, M., M. T. Pilehvar, N. Limsopatham & N. Collier (2018) What's missing in geographical parsing. *Language Resources and Evaluation* 52: 2, 603-623. <*DOI:10.1007/s10579-017-9385-8*>

Gohen, Peter G. (1998) Public space and the geography of the modern city. *Progress in Human Geography* 22: 4, 479-496.

Hausmann, A., T. Toivonen, V. Heikinheimo, H. Tenkanen, R. Slotow & E. Di Minin (2017) Social media reveal that charismatic species are not the main attractor of ecotourists to sub- Saharan protected areas. *Scientific Reports* 7. <*DOI:10.1038/s41598-017-00858-6*>

Hausmann, A., T. Toivonen, R. Slotow, H. Tenkanen, A. Moilanen, V. Heikinheimo & E. Di Minin (2018) Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas. *Conservation Letters* 11: 1. <*DOI: 10.1111/conl.12343*>

Hiippala, T. A. Hausmann, H. Tenkanen & T. Toivonen (2018) Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*

Hiippala, Tuomo (2017) Recognizing military vehicles in social media images using deep learning. *15th IEEE International Conference on Intelligence and Security Informatics - Security and Big Data (ISI)*, 60-65.

Hochmair, H. H., L. Juhász & S. Cvetojevic (2018) Data Quality of Points of Interest in Selected Mapping and Social Media Platforms. Teoksessa Kiefer, P., Huang, H., Van de Weghe, N., Raubal, M (2018) Progress in Location Based Services 2018, pp.293-313, Springer.

Hochman, Nadav & Lev Manovich (2013) Zooming into an Instagram City: Reading the local through social media. *First Monday* 18: 7. *<http://firstmonday.org/ojs/index.php/fm/article/view/4711/3698>* Viitattu 29.03.2017.

Hong, L. & B. D. Davison (2010) Empirical Study of Topic Modeling in Twitter. *Proceedings of the 1st Workshop on Social Media Analytics (SOMA '10)*, 80-88.

Instagram for Developers (2016) Instagram Platform Update *<http://developers.instagram.com/post/133424514006/instagram-platform-update>* Viitattu 20.03.2017

Jauhiainen, T., M. Lui, M. Zampieri, T. Baldwin & K. Linden (2018) Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*. 1-97. *<arXiv:1804.08186>*

Jenkins A, A. Croitoru, A. T. Crooks & A, Stefanidis (2016) Crowdsourcing a Collective Sense of Place. *PLoS ONE* 11: 4. *<DOI:10.1371/journal.pone.0152932>*

Karjalainen, Marketta (2018) Äänestys: Helsingin paras brunssipaikka on Kalliossa. Helsingin Uutiset *<https://www.helsinginuutiset.fi/artikkeli/504661-aanestys-helsingin-paras-brunssipaikka-on-kalliossa-tassa-kaikki-lukijoiden>* Viitattu 20.08.2018

Keipi, T., I. Koiranen, A. Koivula & P. Räsänen (2018) Assessing the social media landscape: Online relational use-purposes and life satisfaction among Finns. *First Monday* 23: 1. *<DOI:10.5210/fm.v23i1.8128>*

Kellerman, Aharon (2014) The Satisfaction of Human Needs in Physical and Virtual Spaces. *The Professional Geographer* 66:4, 538–546

Kellerman, Aharon (2015) Image spaces and the geography of Internet screen-space. *GeoJournal* 81, 501-517.

Kitchin, Robert (1998) Towards geographies of cyberspace. *Progress in Human Geography* 22: 3, 385-406.

Koskenniemi, Timo (2011) Käsitehakemisto: terms and concepts of language technology. <*http://www.ling.helsinki.fi/kit/2004s/terms-en.shtml*> Viitattu 16.11.2018.

Lagus, K., M. Pantzar, M. Ruckenstein ja M. Ylisiurua (2016). Suomi 24 – Muodonantoa aineistolle. *Valtiotieteellisen tiedekunnan julkaisuja* 2016: 10.

Lansley, Guy & Paul A. Longley (2016) The geography of Twitter topics in London. *Computers, Environment and Urban Systems* 58, 85-96.

Larochelle, Hugo & Stanislas Lauly (2012) A Neural Autoregressive Topic Model. *Advances in Neural Information Processing Systems*, 2708-2716.

Lau, J. H., K. Grieser, D. Newman & T. Baldwin (2011) Automatic Labeling of Topic Models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* 1, 1536–1545.

Leppänen, S., Pitkänen-Huhta, A., Nikula, T., Kytölä, S., Törmäkangas, T., Nissinen, K., Kääntä, L., Räisänen, T., Laitinen, M., Pahta, P., Koskela, H., Lähdesmäki, S. & Jousmäki, H. (2011) National survey on the English language in Finland: Uses, meanings and attitudes. *Studies in Variation, Contacts and Change in English* 5

Leung, Louis (2013) Generational differences in content generation in social media: The roles of the gratification sought and of narcissism. *Computers in Human Behavior* 29, 997-1006.

Leung, R., H. Q. Vu & J. Rong (2017) Understanding tourists' photo sharing and visit pattern at non-first tier attractions via geotagged photos. *Information Technology & Tourism* 17: 1, 55–74. <*DOI:10.1007/s40558-017-0078-3*>

Liu, Y., Z. Sui, C. Kang & Y. Gao (2014) Uncovering Patterns of the Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. *PLoS One* 9: 1. *<DOI:10.1371/journal.pone.0086026>* Viitattu 15.04.2018.

Maeda, T. N., M. Yoshida, F. Toriumi & H. Ohashi (2018) Extraction of Tourist Destinations and Comparative Analysis of Preferences Between Foreign Tourists and Domestic Tourists on the Basis of Geotagged Social Media Data. *ISPRS International Journal of Geo-Information* 7: 3, 1-19. *<DOI:10.3390/ijgi7030099>*

Maiya, A. S., J. P. Thompson, F. Loaiza-Lemos & R. M. Rolfe (2013) Exploratory Analysis of Highly Heterogeneous Document Collections. *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 1375–1383. <arXiv:1308.2359>

Malecki, Edward J. (2017) Real people, virtual spaces, and the spaces in between. *Socio-economic Planning Sciences* 58, 3-12.

Manovich, Lev & Agustin Idaco (2017) The Image of a Data City: Studying the Hyperlocal with Social Media. *Architectural Design* 87: 1, 110-117.

Markham, Annette (2012) FABRICATION AS ETHICAL PRACTICE. *Information, Communication & Society* 15: 3, 334-353. *<DOI:10.1080/1369118X.2011.641993>*

Martin, E. M. & N. Schuurman (2017) Area-Based Topic Modelling and Visualization of Social Media for Qualitative GIS. *Annals of the American Association of Geographers* 107: 5, 1028-1039.

May, C., R. Cotterell & B. Van Durme (2016) Analysis of Morphology in Topic Modeling. <arXiv:1608.03995>

Mazarura, J., A. de Waal, F. Kanfer & S. Millard (2014) Topic Modelling for Short Text. *Proceedings of the 2014 PRASA, RobMech and AfLaT International Joint Symposium*, 1-5.

Mehrotra, R., S. Sanner. W. Buntine & L. Xie (2013) Improving LDA topic models for microblogs via tweet pooling and automatic labeling. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 889-892. *<DOI:10.1145/2484028.2484166>*

Miller, H. J. & M. F. Goodchild (2015) Data-driven geography. *GeoJournal* 80: 449 - 461.

Mäkinen, Vesa (2018) Facebookin Zuckerberg todistaa kongressin kauppakomitean edessä ensi viikolla käyttäjätietoskandaalin takia. <*https://www.hs.fi/ulkomaat/art-2000005629098.html*> Viitattu 10.08.2018

Neuhold, G., T. Ollmann, S. R. Bulo & P. Konstchieder (2017) The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. *2017 IEEE International Conference on Computer Vision*, 5000- 5009.

Newton, Casey (2016) Instagram is getting rid of photo maps. <*https://www.theverge.com/2016/9/6/12817340/instagram-photo-map-removals*> Viitattu 20.03.2018

Papacharissi, Zizi (2002) The virtual sphere: The internet as a public sphere. *New Media & Society* 4: 1, 9–27. <DOI:10.1177/14614440222226244>

Peukert, H. (2013). Measuring linguistic diversity in urban ecosystems. *Teoksessa* Duarte, J. and Gogolin, I. (toim.) *Linguistic Superdiversity in Urban Areas: Research Approaches*, 75–93. Benjamins Publishing, Amsterdam.

Places (2015) Mission Control: History of the Urban Dashboard. <*https://placesjournal.org/article/mission-control-a-history-of-the-urban-dashboard/*> Viitattu 13.04.2018.

Ramage, D., S. Dumais & D. Liebling (2010) Characterizing Microblogs with Topic Models. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 130-137.

Redi, M., D. Crocket, L. Manovich & S. Osindero (2016) What makes photo cultures different? *Proceedings of the 2016 ACM on Multimedia Conference*, 287-291.

Rehurek, Radim (2018) Gensim Tutorial 2: Topics and Transformations <*https://radimrehurek.com/gensim/tut2.html*> Viitattu. 25.03.2018.

Rose, Gillian (2017) Posthuman Agency in the Digitally Mediated City: Exteriorization, Individuation, Reinvention. *Annals of the American Association of Geographers* 107:4, 779-793.

Rosen-Zvi, M., Griffiths, T.L., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence. *<arXiv:1207.4169>*

Sheldon, Pavica & Katherine Bryant (2016) Instagram: Motives for its use and relationship to narcissism and contextual age. *Computers & Human Behavior* 58, 89-97. *<DOI:10.1016/j.chb.2015.12.059>*

Silfverberg, M., T. Ruokolainen, K. Lindén and M. Kurimo (2016). FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish. *Language Resources and Evaluation* 50: 4, 863-878.

Silva, T. H., P. O. S. Vaz de Melo, J. M. Almeida & A. A. F. Loureiro (2014) Large-scale Study of City Dynamics and Urban Social Behavior Using Participatory Sensing. *IEEE Wireless Communications* 21: 1, 42-51.

Stadissa (2018) Kallion brunssit. Inessiivi Media Oy *<http://www.stadissa.fi/paikat/649/kallion-brunssit>* Viitattu 20.08.2018

Srinivasan, S., S. Bhattacharya & R. Chakraborty (2012) Segmenting Web-Domains and Hashtags using Length Specific Models. *21st ACM International Conference on Information and Knowledge Management*, 1113-1122.

Statista (2018a) Global social media ranked by number of users. *<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>* Viitattu 05.04.2018

Statista (2018b) IoT: number of connected devices worldwide 2015-2025. *<https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/>* Viitattu 13.09.2018

Statista (2018c) Smartphone shipments worldwide 2009-2018. <*https://www.statista.com/statistics/271491/worldwide-shipments-of-smartphones-since-2009/*> Viitattu 13.09.2018

Statista (2018d) Number of social media users worldwide 2010-2021. <*https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/*> Viitattu 13.09.2018

Stefanidis, A., A. Crooks & J. Radzikowski (2013) Harvesting ambient geospatial information from social media feeds. *GeoJournal 78*, 319-338. <*DOI:10.1007/s10708-011-9438*>

SYKE (2016) Yhdyskuntarakenteen seurannan aineistot. Suomen Ympäristökeskus, Helsinki.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll & M. Stede (2011) Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* 37: 2, 267-307.

Tenkanen, Henrikki (2017) Capturing time in space - Dynamic analysis of accessibility and mobility to support spatial planning using open data and tools. 182 s. Väitöskirja, Geotieteiden ja maantieteen laitos, Helsingin yliopisto. <*http://urn.fi/URN:ISBN:978-951-51-2935-9*> Viitattu 14.11.2018

Tenkanen, H., E. Di Minin, V. Heikinheimo, A. Hausmann, M. Herbst, L. Kajala & T. Toivonen (2017) Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports* 7. <*DOI:10.1038/s41598-017-18007-4*> Viitattu 10.05.2018.

Terras, Melissa (2011) Quantifying Digital Humanities <*http://www.ucl.ac.uk/infostudies/melissa-terras/DigitalHumanitiesInfographic.pdf*> UCL Center for Digital Humanities. Viitattu 29.10.2017

Tieteen termipankki (2018) Kielitiede: Polysemia. <*http://www.tieteentermipankki.fi/wiki/Kielitiede:polysemia*> Viitattu 10.10.2018.

Townsend, Anthony M. (2013) *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. 1. p. 400 s. W. W. Norton & company, New York, Yhdysvallat.

Tranos, Emmanouil & Peter Nijkamp (2013) The Death Of Distance Revisited: Cyber-place, Physical And Relational Proximities. *Journal Of Regional Science* 53: 5. 855–873. *<DOI:10.1111/jors.12021>*

Urry, John & Jonas Larsen (2011) *The Tourist Gaze 3.0*. 3. p. 282 s. SAGE Publications Ltd, Lontoo, Iso-Britannia.

Varol, O., E. Ferrara, C. A. Davis, F. Menczer & A. Flammini (2017) Online Human-Bot Interactions: Detection, Estimation and Characterization. *International AAAI Conference on Web and Social Media 2017*. *<arXiv:1703.03107>*

Vu, H. Q., G. Li, R. Law & Y. Zhang (2017) Tourist Activity Analysis by Leveraging Mobile Social Media Data. *Journal of Travel Research* 57: 7, 883–898. *<DOI:10.1177/0047287517722232>*

Wagner, Kurt (2016) Instagram is still growing quickly and now has 600 million users. *<https://www.recode.net/2016/12/15/13971624/instagram-600-million-users-growth>* Viitattu 20.03.2017

Wong, J. C. (2018) Mark Zuckerberg faces tough questions in two-day congressional testimony *<https://www.theguardian.com/technology/live/2018/apr/11/mark-zuckerberg-testimony-live-updates-house-congress-cambridge-analytica>* Viitattu 13.04.2018

Wylie, Christopher (2018) Why I broke the Facebook data story - and what should happen now. *<https://www.theguardian.com/uk-news/2018/apr/07/christopher-wylie-why-i-broke-the-facebook-data-story-and-what-should-happen-now>* Viitattu 13.04.2018

Xu, Z., Y. Liu, N. Y. Yen, L. Mei, X. Luo, X. Wei & C. Hu (2015) Crowdsourcing based Description of Urban Emergency Events using Social Media Big Data. *IEEE Transactions on Cloud Computing* 99. *<DOI:10.1109/TCC.2016.2517638>*

Yang, L., L. Wu, Y. Liu & C. Kang (2017) Quantifying Tourist Behavior Patterns by Travel Motifs and Geo-Tagged Photos from Flickr. *ISPRS International Journal of Geo-Information* 6: 11, 1-18. *<DOI:10.3390/ijgi6110345>*

Zhao, W. X., J. Jing, W. Jianshu, H. Jing, L. Ee-Peng, Y. Hongfei & L. Xiaoming (2011) Comparing Twitter and Traditional Media Using Topic Models. *Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR*, 1-14.

Zhou, B., L. Liu, A. Olivia & A. Torralba (2014) Recognizing City Identity via Attribute Analysis of Geo-tagged Images. *Teoksessa* Fleet D., Pajdla T., Schiele B., Tuytelaars T. (toim) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science 8691. Springer, 519-534.

Zhou, Xiaolu & Liang Zhang (2016) Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartography and Geographic Information Science* 43: 5, 393-404.

Zook, Matthew & Mark Graham (2007) Mapping DigiPlace: geocoded Internet data and the representation of place. *ENVIRONMENT AND PLANNING B-PLANNING & DESIGN* 34: 3, 466-482 *<DOI:10.1068/b3311>*

Zook, Matthew & Mark Graham (2017) Hacking Code/Space: Confounding the code of global capitalism. *Transactions of the Institute of British Geographers* 43: 3, 390-404. *<DOI:10.1111/tran.12228>*