Aalto-yliopisto
Insinööritieteiden korkeakoulu

Geokoodausmenetelmät ja niiden käyttö sosiaalisen median julkaisujen paikantamisessa

Diplomityö
Rakennetun ympäristön laitos
Insinööritieteiden korkeakoulu
Aalto-yliopisto

Espoossa 29 toukokuuta 2017

Tekniikan kandidaatti Hanna Grannabba

Valvoja: Professori Kirsi Virrantaus
Ohjaaja: DI Thomas Nynäs

Aalto-yliopisto
Insinööritieteiden korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Diplomityön tiivistelmä

Tekijä Hanna Grannabba

Työn nimi Geokoodausmenetelmät ja niiden käyttö sosiaalisen median julkaisujen paikantamisessa

Koulutusohjelma Geomatiikka

Pääaine Geoinformatiikka **Koodi** M3002

Työn valvoja Professori Kirsi Virrantaus

Työn ohjaaja(t) Thomas Nynäs

Päivämäärä 29.05.2017 **Sivumäärä** 56 **Kieli** Suomi

Tiivistelmä

Monet aineistot sisältävät osoitteita ja paikannimiä. Jotta aineistoja voidaan hyödyntää paikkatietoanalyyseissä, ne on georeferoitava eli niille on saatava koordinaatit. Geokoodaus on prosessi, jossa osoitteelle tai paikannimelle pyritään selvittämään sijainti vertailuaineiston avulla. Sosiaalinen media on nykyisin tärkeä osa ihmisten elämää, mikä on synnyttänyt tarpeen useilla eri aloilla pystyä selvittämään, mistä paikasta sosiaalisen median julkaisut on tehty tai mitä paikkaa ne koskevat. Tämän diplomityön tarkoituksena oli tutkia geokoodausta, siihen käytettäviä menetelmiä sekä niiden ominaisuuksia.

Työn teoreettisessa osiossa käydään läpi geokoodausprosessin kulku ja miten osoitteiden geokoodaus tapahtuu erilaisia vertailuaineistoja käyttämällä. Sen jälkeen perehdytään twiittien geokoodaukseen ja siihen mitä lisähaasteita se tuottaa verrattuna osoitteiden geokoodaukseen. Työn empiirisessä osiossa toteutettiin työkaluja twiittien geokoodausta varten sekä testattiin niiden toimivuutta käytännössä.

Työssä havaittiin, että geokoodaus on monivaiheinen prosessi, jossa on tehtävä kompromisseja tulosten tarkkuuden, kattavuuden ja ajankäytön välillä. Tutkimuksessa verrattiin kolmea erilaista yhdistämisalgoritmia, Levenshtein distance, Longest common subsequence ja n-gram, joilla samankaltaiset merkkijonot voidaan yhdistää toisiinsa. Näistä algoritmeista n-grammeihin perustuva vertailu tuotti tarkimman tuloksen. Suurimmaksi haasteeksi havaittiin paikannimien erottaminen tavallisten sanojen joukosta, eli geoparsing. Monet tavalliset sanat esiintyvät myös paikanniminä joissain päin maailmaa, mikä aiheuttaa virheellisiä paikannuksia, ellei niitä pystytä havaitsemaan.

Avainsanat geokoodaus, georeferointi, twitter, osoite, paikannimi

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author	Hanna Grannabba
Title of thesis	Geocoding methods and their usage in locating social media posts
Degree programme	Degree Programme in Geomatics
Major	Geoinformatics
Code	M3002
Thesis supervisor	Professor Kirsi Virrantaus
Thesis advisor(s)	Thomas Nynäs
Date	29.05.2017
Number of pages	56
Language	Finnish

Abstract

Many documents contain addresses and place names. In order to make spatial analysis for these documents they need to be georeferenced. Which means they need coordinates. Geocoding is a process where addresses and place names are given coordinates based on a reference dataset. Social media is an important part of peoples' life nowadays and there is an increasing need for knowing where the posts are sent from or what place they refer to. The purpose of this masters' thesis was to examine geocoding, the methods used for it and their features.

The theoretical part of the study presents the geocoding process and how address geocoding is done with different types of reference datasets. In addition, geocoding of tweets is examined, and what additional challenges it does have compared to address geocoding. In the practical part of the study tools for geocoding tweets where implemented and tested in practice.

It was noticed that geocoding consists of many phases and it is necessary to make compromises between accuracy, completeness and execution time. Three different feature matching algorithms, Levenshtein distance, Longest common subsequence and n-gram, where tested. With feature matching strings that are approximately similar can be combined. Of these three the one based on n-grams gave the most accurate results. The biggest challenge appears to be recognizing place names among all other words, called geoparsing. Several normal words occur as place names on different places in the world. If these can't be distinguished they will cause false matches in the geocoding results.

Keywords geocoding, georeferencing, twitter, address, place name

Alkusanat

Idea tämän työn aiheeseen tuli InPlace Solutionsilta ja se muovautui matkan varrella yhteistyössä heidän kanssaan. Haluan kiittää InPlace Solutionsia työn tekemisen mahdollistamisesta. Kiitän kaikkia työkavereita ja erityisesti ohjaajaani Thomas Nynäsiä ideoista ja palautteesta. Lisäksi kiitän valvojaani Kirsi Virrantausta hyvistä neuvoista sekä perhettäni tuesta koko opiskeluaikana.

Espoo 29.5.2017

Hanna Grannabba

Sisällysluettelo

1 Johdanto

1.1 Työn tausta ja tavoitteet

Paikannimet ja osoitteet ovat tapa jolla ihmiset muodostavat käsityksen itseään ympäröivästä maailmasta. Jotta niitä voidaan esittää kartalla tai hyödyntää paikkatietoanalyyseissä on niille saatava koordinaatit. Geokoodaus on prosessi, jolla osoitteen tai paikannimen koordinaatit pyritään selvittämään vertailuaineistoon vertaamalla.

Suuri osa valtiollisista ja kaupallisista aineistoista sisältää osoitetietoja ja geokoodauksesta on tullut tärkeää, sillä se mahdollistaa paikkatietoanalyysien tekemisen näille aineistoille. (Christen ym., 2004) Alkuaikoina geokoodausta tekivät ammattilaiset, jotka tunsivat prosessin ja siihen liittyvät epävarmuudet. Internetin ja web-palveluiden myötä on tullut mahdolliseksi melkein kenelle tahansa muuttaa osoite koordinaateiksi käyttämällä siihen tarkoitettua palvelua, joko täysin ilmaiseksi tai melko halvalla. (Roongpiboonsopit & Karimi, 2010) Geokoodaukseen sisältyy kuitenkin edelleen lukuisia epävarmuustekijöitä ja siksi tutkijoiden tulee olla tietoisia siitä, miten geokoodauksessa tehdyt valinnat vaikuttavat geokoodauksen tulokseen ja sen myötä koko tutkimuksen tulokseen. (Cayo ja Talbot, 2003)

Sosiaalisen median käytön lisääntyessä on tullut tarve selvittää, mistä julkaisut on lähetetty tai mitä paikkaa ne koskevat. Paikannettuja julkaisuja voidaan hyödyntää monella alalla aina luonnonkatastrofien pelastustöistä mainontaan. Osaan julkaisuista on liitettynä käyttäjän puhelimen koordinaatit, mutta useimmiten sijainti joudutaan selvittämään julkaisun tekstin tai muiden metatietojen perusteella. (Han ym., 2014) Tekstissä voi olla mainittuna paikka, jonka perusteella julkaisun sijainti voidaan selvittää. Paikan erottaminen sosiaalisen median tekstistä on haastavampaa kuin esimerkiksi uutisartikkelista. Sosiaalisessa mediassa julkaistavat tekstit ovat usein lyhyitä ja puhekielen sekä murresanojen käyttö on yleistä. Koska julkaisut tehdään spontaanisti ja usein mobiililaitteella, sisältävät ne usein myös kirjoitusvirheitä.

Muutamissa tutkimuksissa on myös hyödynnetty paikannimien lisäksi muita sanoja twiittien paikantamiseen (Zhang & Gelernter, 2014). Tällaisia sanoja voivat olla esimerkiksi tapahtumat ja urheilujoukkueet. Useimmat twitterin geokoodausta koskevat aikaisemmat tutkimukset käsittelevät käyttäjän kotipaikan selvittämistä, käyttämällä useita twiittejä samalta käyttäjältä. Tässä tutkimuksessa sen sijaan pyritään selvittämään tapa yksittäisen twiitin paikantamiseen.

Työn tarkoituksena on tutkia geokoodausmenetelmien ominaisuuksia. Tavoitteena on selvittää mitä menetelmiä twiittien, ja siten myös muiden sosiaalisen median julkaisujen, geokoodaukseen voidaan käyttää. Prosessin selvittämiseksi käydään ensin läpi osoitteiden geokoodaus, jonka jälkeen keskitytään twiittien geokoodaukseen.

Työn tutkimuskysymykset ovat:

- Mitä geokoodaus on?
- Millä menetelmillä sitä voidaan tehdä?
- Millaisia ominaisuuksia näillä menetelmillä on?

Työn aihepiiri on laaja ja siksi eri osa-alueet rajataan tarkasteltaviksi melko yleisellä tasolla, eikä esimerkiksi algoritmien laskenta- ja toteutustapoihin syvennytä tarkemmin. Tutkimus

tehtiin yhteistyössä InPlace Solutionsin kanssa. Heidän toimintansa vaatii aineistojen paikantamista tavalla, johon käytettävissä olevat kaupalliset geokoodauspalvelut eivät tällä hetkellä pysty. Paikantamisen lisäksi heille on tärkeää, saada tietää millä menetelmillä ja ratkaisuilla geokoodauksen tulokseen päädyttiin. Työssä pyritään tuomaan esille mitä kaikkea geokoodauksessa on otettava huomioon, eikä varsinaisesti tuottamaan valmista optimaalista järjestelmää.

1.2 Tutkimusmenetelmät

Työn teoreettinen osa muodostuu kirjallisuustutkimuksesta, jolla luodaan pohja työn empiiriselle osiolle. Kirjallisuustutkimuksessa perehdytään geokoodausta koskeviin aikaisemmin tehtyihin tieteellisiin tutkimuksiin. Geokoodausta ovat tutkineet mm. Peter Christensen, Paul A. Zandbergen ja Dan Goldberg. Twittergeokoodaukseen ovat aikaisemmin perehtyneet mm. Han ym. (2014), Zhang ja Gelernter (2014) sekä Sidkar ja Gambäck (2016). Työn empiirisessä osiossa kehitetään Esrin ArcMapiin työkalupaketti twiittien geokoodausta varten. Työkalupakettiin toteutetaan työkaluja, jotka käyttävät erilaisia menetelmiä ja näitä testataan käytännössä. Testaus suoritetaan twitteristä kerätyllä aineistolla ja tulosten paikkansapitävyyttä arvioidaan käymällä ne manuaalisesti läpi. Tarvittava aineisto tulee InPlace Solutionsilta.

1.3 Työn rakenne

Työn toisessa kappaleessa käydään läpi geokoodausprosessi sekä aikaisempaa tutkimusta. Ensin kuvaillaan geokoodauksen perustapaus, eli osoitteiden paikantaminen, ja minkälaisia vaihtoehtoja sen suorittamiseen on olemassa. Sen jälkeen keskitytään twiittien geokoodaukseen ja niiden tuomiin lisähaasteisiin. Kolmas kappale käsittelee työn empiiristä osiota. Siinä käydään läpi geokoodaustyökalujen toteutus ja niiden toimintaperiaatteet sekä suoritetut testit. Neljännessä kappaleessa tarkastellaan testiaineistoilla saatuja tuloksia sekä pohditaan niiden syitä. Kappaleessa viisi esitetään ehdotuksia järjestelmän jatkokehitystä varten ja viimeisessä, eli kuudennessa, kappaleessa esitellään työn johtopäätökset.

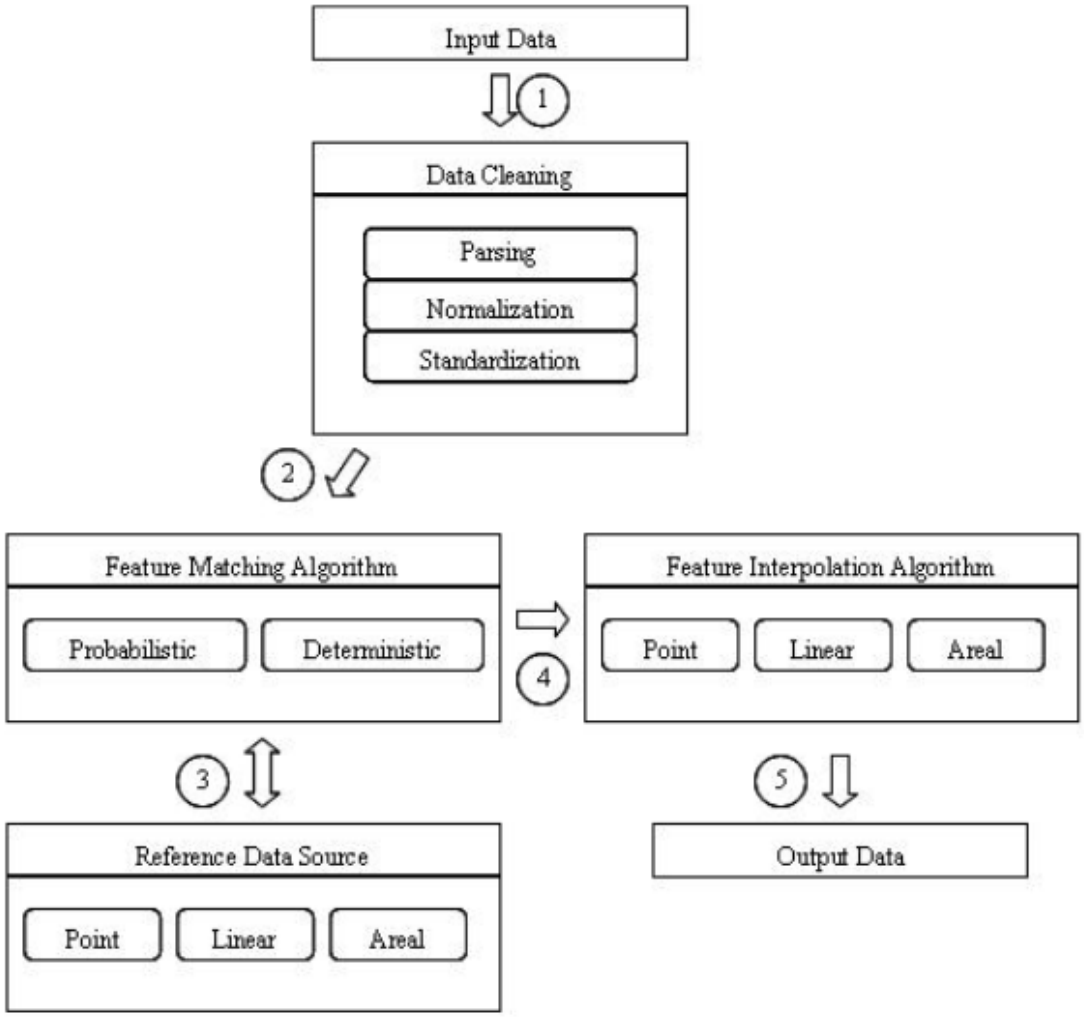
2 Geokoodauksen perusteet

2.1 Prosessi

Geokoodauksella pyritään selvittämään verbaalisen paikan ilmaisun sijainti yhdistämällä se valmiiksi georeferoituun vertailuaineistoon, jotta sille saadaan koordinaatit. Geokoodattua tietoa voidaan visualisoida kartalla tai yhdistää muuhun paikkatietoon sekä analysoida. (Christen ym., 2004). Hyvin yleistetysti prosessin voidaan sanoa koostuvan kolmesta osasta, jotka ovat syöte, geokoodaus ja tulos. Syöte, eli geokoodattava tieto voi olla esimerkiksi osoite, paikannimi tai muunlainen kuvaus paikasta, jolle halutaan kohde kartalla. Geokoodaus sisältää itsessään useita vaiheita. Osoitetta geokoodattaessa osoite jaetaan ensin komponentteihin, kuten kadun nimi, kadun tyyppi, postinumero ja kaupunki. Tämän jälkeen nimet normalisoidaan, jotta päästään eroon eriävistä kirjoitusasuista. Osoitteita verrataan vertailuaineistoon ja jokaiselle osumalle annetaan arvo, sen mukaan miten hyvin se vastaa etsittyä osoitetta. Lopulta osumista valitaan osoitetta parhaiten vastaava, joka on prosessin tulos. (Zandbergen, 2008; Goldberg, 2008) Osa tutkijoista erottaa geokoodauksen ja georeferoinnin toisistaan, jolloin geokoodauksella tarkoitetaan osoitteen muuttamista koordinaateiksi ja georeferoinnilla paikannimen muuttamista koordinaateiksi. Käytännössä geokoodauksella kuitenkin viitataan yleensä molempiin. (Goldberg, 2013)

Eri tahojen keräämien osoitetietojen ulkomuodot voivat olla hyvin vaihtelevia ja siksi niistä on tehtävä vertailukelpoisia toisiinsa ja vertailuaineistoon nähden. Tarkkojen geokoodaustulosten saamiseksi on tärkeää, että sekä käyttäjän antama syöte, eli geokoodattava osoite tai paikannimi, että vertailuaineisto on luokiteltu ja normalisoitu samalla tavalla. Goldbergin (2008) mukaan syötteen ”puhtaus” saattaa olla merkittävin vaikuttava tekijä siihen onnistuuko geokoodaus vai ei.

Kuva 1 esittää Goldbergin (2008) näkemyksen geokoodausjärjestelmän rakenteesta. Aineiston siivoamisvaihe koostuu useista vaiheista ja kaikki tutkijat eivät erottele ja nimitä niitä samalla tavalla kuten Goldberg. Datan siivoaminen alkaa osoitteen luokittelulla, jossa osoite jaetaan osiin ja osille annetaan tiettyjen sääntöjen mukaan niin kutsuttu match key, eli tieto siitä, onko kyseessä esim. tie, postinumero tai jokin tuntematon osa osoitetta. Normalisoinnilla muunnetaan mahdolliset erilaiset kirjoitusasut tai versiot samasta osoitteesta samanlaisiksi tiettyjen sääntöjen mukaan. Tämä tapahtuu esimerkiksi vaihtamalla kaikki kirjaimet pieniksi ja korvaamalla osoitteissa esiintyvät samaa tarkoittavat lyhenteet yhdellä määrätyllä lyhenteellä. (Christen ym., 2004; Goldbergin, 2008; Zandbergen, 2008) Esimerkiksi kaikki tiennimet joissa esiintyy ”street” voidaan muuttaa muotoon ”st”. Osoitteen normalisointia käydään tarkemmin läpi kappaleessa 2.1.1.



Kuva 1. Geokoodausprosessi. Goldberg (2008)

Vertailuaineisto sisältää tiedossa olevat maantieteelliset kohteet, joko piste, viiva tai aluemuodossa, ja niiden perusteella pystytään määrittämään haetun osoitteen sijainti. Vertailuaineiston lähteenä voi olla esimerkiksi alueellinen tai kansallinen organisaatio. (Goldberg, 2008)

Algoritmit ovat pääasiallinen laskennan suorittava osa ja ne vaihtelevat järjestelmästä toiseen. Yhteistä niille kuitenkin on, että ne koostuvat kohteen yhdistämisalgoritmista (feature matching algorithm) sekä interpolointialgoritmista. Algoritmit voidaan toteuttaa monella eri tapaa, etenkin jos syötteen tai vertailuaineiston luonne tiedetään etukäteen. Kohteen yhdistämisalgoritmi yrittää löytää yhden tai useamman kohteen vertailuaineistosta, joka vastaa etsittyä osoitetta. Algoritmi on riippuvainen vertailuaineiston tyypistä sekä sen kohteiden attribuuteista ja se voi olla joko interaktiivinen, eli vaatia käyttäjää tekemään valintoja algoritmin itsensä epäonnistuessa, tai se voi olla täysin automaattinen. Yhdistämisprosessi voi olla joko deterministinen tai todennäköisyyspohjainen. Deterministinen järjestelmä joko löytää osuman tai ei. Hyvänä puolena on, että tällainen järjestelmä antaa aina saman tuloksen tiettyä syötettä käytettäessä. Todennäköisyyspohjainen järjestelmä palauttaa joukon osumia, sekä tiedon siitä millä todennäköisyydellä mikäkin tulos on oikea. Tulos saattaa kuitenkin vaihdella vertailuaineistosta riippuen. (Goldberg, 2008; Goldberg, 2013)

Myös interpolointialgoritmi riippuu vahvasti vertailuaineiston tyypistä. Jos kyseessä on pisteaineisto, on tulos suoraan yhdistämisalgoritmin palauttama kohde eikä interpolointia tarvita. Viiva tai aluemuodossa olevaa aineistoa käytettäessä on suoritettava interpolointi, jotta saadaan tietää missä syötteen osoite sijaitsee suhteessa vertailuaineistosta valittuun kohteeseen. (Goldberg, 2008) Interpolointimenetelmiä käsitellään tarkemmin eri vertailuaineistoihin perustuvien menetelmien kohdalla.

Geokoodauksen tuloksena saadaan useimmiten yhden pisteen koordinaatit. Koordinaattien lisäksi tuloksen on Goldbergin (2008) mukaan hyvä sisältää metatietoa geokoodauksen kulusta ja sen aikana tehdyistä valinnoista. Metatiedot voivat esimerkiksi kertoa syötteen normalisoinnista, vertailuaineistosta sekä yhdistämiseen ja interpolointiin käytetyistä algoritmeista. Ainoastaan yhden pisteen palauttaminen aiheuttaa ongelmia tapauksissa, joissa kaksi eri sijaintia ovat yhtä todennäköisesti oikein. Goldberg (2013) käyttää esimerkkinä tilannetta, jossa on haettu osoitetta ”123 Main Street”. Tällöin sekä ”123 N Main Street” että ”123 S Main Street” ovat yhtä todennäköisesti se mitä käyttäjä todellisuudessa haki. On olemassa eri vaihtoehtoja sille, miten geokoodausjärjestelmä käsittelee tällaista tilannetta. Jotkut ohjelmat arpovat kumpi palautetaan, toiset palauttavat sen kumpaa on haettu useammin ja osa ilmoittaa epäonnistuneensa eikä palauta kumpaakaan. Yksi ratkaisu tähän ongelmaan on palauttaa kaikki vaihtoehdot, jolloin valitsemisen vastuu siirtyy käyttäjälle. (Goldberg, 2013)

2.1.1 Osoitteen normalisointi

Osoitteen jaottelu eri luokkiin on tärkeä vaihe siivoamisen kannalta, sillä ilman oikeanlaista luokittelua on mahdotonta yhdistää osoitetta vertailuaineistoon. Jaotteluun voidaan käyttää monia eri menetelmiä, joista tässä esitellään pari. Korvaamispohjainen (substitution-based normalization) on suosittu metodi johtuen sen helppokäyttöisyydestä. Sitä olisikin Goldbergin (2008) mukaan hyvä käyttää ensimmäisenä vaiheena ja etenkin jos muita metodeja ei ole käytettävissä. Osoite jaetaan välilyöntien perusteella osiin ja osia vertaillaan yksi kerrallaan tiettyihin sääntöihin. Esimerkkinä osoite “3620 Vermont Ave, RM444, Los Angeles, CA 90089”. Ensimmäisen osan tunnistetaan olevan numero, jonka jälkeen tarkastellaan, onko seuraava osa Yhdysvalloissa usein käytetty ilmansuuntaa kuvaava sana tai lyhenne, esim. n tai north. Koska näin ei ole, on toisen osan oltava tiennimi. ”Ave” puolestaan sopii kategoriaan tien tyyppi ja luokitellaan siihen ja näin jatketaan koko osoitteen läpi. Tämä menetelmä aiheuttaa ongelmia, jos tien nimessä esiintyykin sanoja, jotka on luokiteltu kuvaamaan esimerkiksi tien tyyppiä. (Goldberg, 2008)

Kontekstiin perustuvan normalisoinnin (context-based normalization) suurin hyöty on sen kyky järjestää osoitteen osat uudelleen, samalla se on kuitenkin hankalampi toteuttaa kuin korvaamispohjainen normalisointi ja siksi harvemmin käytetty. Ensimmäisessä vaiheessa osoitteesta poistetaan kaikki erikoismerkit ja ylimääräiset välilyönnit ja erotellaan sanat yksittäisillä välilyönneillä. Tämän jälkeen muutetaan kaikki merkit joko isoiksi tai pieniksi kirjaimiksi. Seuraavassa vaiheessa merkkijonoille määrätään tyyppi, sen mukaan onko kyseessä numero, kirjaimia vai yhdistelmä molempia. Viimeisessä vaiheessa merkkijonot sijoitetaan parse tree nimiseen puutietorakenteeseen kielioppiin perustuen. Kielioppi on tässä tapauksessa joukko sääntöjä sille, mistä osista osoite voi koostua. Osoitteessa voi olla katuosoiteosa ja paikkakuntaosa. Katuosoiteosa puolestaan koostuu talonnumerosta ja kadunnimestä jne. Menetelmän hankaluus johtuu siitä, että yksi merkkijono voi sääntöjen puolesta sopia useampaan kuin yhteen paikkaan, jolloin saadaan aikaan näennäisesti kelvollinen osoite, joka kuitenkin on väärä. Tästä syystä on asetettava rajoituksia sille, miten kunkin

tyyppisiä merkkijonoja voidaan siirtää ja mahdollisesti suorittaa iteratiivinen prosessi, joka aloittaa alkuperäisen osoitteen järjestyksellä. (Goldberg, 2008)

2.1.2 Vertailuaineisto

Vertailuaineisto on maantieteellinen nimihakemisto, jota kutsutaan englanniksi nimellä gazetteer, ja se koostuu paikannimistä ja niitä vastaavista maantieteellisistä kohteista sekä mahdollisesti muista kuhunkin paikkaan liittyvistä lisätiedoista. Vertailuaineisto tallennetaan usein puutietorakenteeseen, hajautustauluun tai SQL tietokantaan (Leidner & Lieberman, 2011). Esimerkiksi Freire ym. (2011) käyttävät vertailuaineiston tietorakenteena B-puuta. Yleisiä vertailuaineistoja joita käytetään paikannimiä geokoodattaessa ovat GeoNames, NGAn GNS (GEOnet Names Server) ja USGSn GNIS (Geographic Names Information System). (Leidner & Lieberman, 2011)

Kun halutaan geokoodata eri maissa sijaitsevia osoitteita, on vertailuaineistojen yhteensovittaminen yksi kriittinen tekijä. Lennert (2015) listaa työssään muutamia kysymyksiä, jotka vaikuttavat siihen miten yksinkertaista aineostoa on käyttää geokoodaukseen. Onko saman postinumeron alueella useita samannimisiä teitä? Käytetäänkö paikannimiä kuten esim. rakennusten nimiä postiosoitteiden asemasta? Löytyykö aineistosta vanhoja osoitteita, esimerkiksi jos tien nimi on muuttunut?

Avoimen käyttäjien tuottaman datan kehittymisen myötä on tullut mahdolliseksi kehittää geokoodausmenetelmiä, jotka eivät vaadi kalliiden aineistojen ostamista maanmittaustoimistoilta tai yksityisiltä yrityksiltä. Amelunxen (2010) on kehittänyt geokoodaussovelluksen joka perustuu OpenStreetMapin (OSM) osoitetietoihin. Kompensoidakseen puutteellisia talonnumerotietoja hän käyttää todennäköisyyspohjaisia interpolointimenetelmiä. Kokeilun tuloksista selviää, että koealueella Saksan Northrhine-Westfaliassa osoitteet joille löytyi täydellinen vastine OSM:stä onnistuttiin geokoodaamaan 11 metrin tarkkuudella, mikä on poikkeuksellisen tarkka tulos aikaisempiin tutkimuksiin verrattuna. Tähän kategoriaan kuuluu kuitenkin vain 5 % etsityistä osoitteista. Vähintään kadun tarkkuudella pystyttiin puolestaan paikantamaan 83 %. Koska Ratcliffen mukaan osumaosuuden tulee olla yli 85 %, jotta geokoodauksen tuloksen perusteella tehdyistä paikkatietoanalyyseistä voidaan luotettavasti havaita ilmiöitä, ei OSM:n perusteella geokoodattua aineistoa ole vielä tutkimuksen teko vaiheessa mahdollista käyttää paikkatietoanalyyseihin. Talonnumerojen määrää aineistossa on kuitenkin jatkuvassa kasvussa, joten tulos on siihen nähden lupaava. (Amelunxen, 2010)

2.2 Yhdistämisalgoritmit

On olemassa lukuisia eri yhdistämisalgoritmeja, joilla kaikilla on sekä hyvät, että huonot puolensa. Yksinkertaistettuna algoritmin tarkoitus on valita vertailuaineistosta kohde joka vastaa syötettyä osoitetta (Goldberg, 2008) ja siksi sitä voidaan pitää tiedostojen yhdistämisen (record linkage) erityistapauksena. Tiedostojen yhdistämisessä tarkoituksena on löytää kaksi oliota, jotka viittaavat samaan tosimaailman olioon, eli tässä tapauksessa paikkaan. (Ranzijn, 2013) Yhdistämisalgoritmin muoto riippuu vertausaineiston tallennustavasta. Vertailuaineiston ollessa relaatiotietokanta toteutetaan yhdistäminen SQL-kyselyillä (Structured Query Language) (Goldberg, 2008). Koska tämän työn tarkoitus on löytää sopiva menetelmä ison aineiston geokoodaukseen, ei interaktiivinen yhdistämisalgoritmi voi tulla kyseeseen ja siksi seuraavaksi tarkastellaan pelkästään automaattisia algoritmeja.

Yhdistämisalgoritmit voidaan luokitella deterministisiin ja todennäköisyyspohjaisiin menetelmiin. Deterministinen menetelmä koostuu joukosta binäärioperaatioita, jotka toteutetaan

tietyssä järjestyksessä ja niiden perusteella kohteelle joko löytyy tai ei löydy osumaa. Tämä tekee menetelmän toteutuksen yksinkertaiseksi, mutta samalla on olemassa riski, ettei se onnistu yhdistämään kohteita. Jos algoritmin tehtävä on etsiä kohde, jonka kaikki attribuutit vastaavat syötteen attribuutteja, mutta vertailuaineistossa onkin olemassa enemmän attribuutteja, niin vastaavaa kohdetta ei löydy. Tästä syystä yhdistelysääntöjä pääsääntöisesti kevennetään siten, että kaikkien attribuuttien ei tarvitse täsmätä vertailuaineistoon. Goldbergin (2008) mielestä yhdistäminen kannattaisi aloittaa deterministisellä menetelmällä ja vertailusäännöt kannattaa ryhmitellä siten, että ensin käytetään tiukkoja kriteerejä, joita lähdetään askel askeleelta keventämään. (Goldberg, 2008)

Todennäköisyyspohjainen yhdistäminen (probabilistic matching) pohjautuu todennäköisyyteen ja päätösteoriaan. Sen matemaattiset mallit saattavat olla monimutkaisia, mutta pääperiaate on kuitenkin yksinkertainen. (Goldberg, 2008) Todennäköisyyspohjainen yhdistäminen on käytössä useissa hakukoneissa, joissa käyttäjälle ehdotetaan jotain toista hakusanaa, jolla löytyy enemmän tuloksia. Sillä pyritään korjaamaan pienet käyttäjän tekemät kirjoitusvirheet. (Nikita, 2011) Todennäköisyyspohjaisissa menetelmissä on aina päätettävä mitä raja-arvoa käytetään määrittämään osuma. Goldberg (2008) suosittelee raja-arvoksi 95 % ja todennäköisyyspohjaisen menetelmän käyttöä silloin, kun deterministinen menetelmä ei tuota tulosta. Mikäli raja-arvo asetetaan korkealle (99 %) pienenee virheellisten osumien riski, mutta samalla osumia saatetaan karsia pois, vaikka niiden kuuluisi olla mukana. Toisaalta jos raja-arvo on kovin matala (esim. 75 %) saatetaan osumiksi virheellisesti laskea sellaisia, jotka eivät oikeasti täsmää. Kumpi on parempi vaihtoehto, riippuu geokoodauksen käyttötarkoituksesta sekä siitä tullaanko tulosta vielä käsittelemään vai käytetäänkö sitä sellaisenaan. (Goldberg, 2013) Goldbergin esittämät luvut eivät päde kaikkiin tapauksiin, vaan sopiva raja riippuu aina geokoodattavasta aineistosta ja sen käyttötarkoituksesta. Jos aineisto on kohtalaisen pieni ja hankala geokoodattava saattaa 60 % olla aivan riittävä raja-arvoksi.

Charif ym. (2010) hyödyntävät opetusmenetelmää yhdistämisprosessin nopeuttamiseksi. Siinä jokainen virhe, joka onnistutaan korjaamaan, lisätään uutena kyseisen sanan mahdollisena kirjoitusasuna tietokantaan, jolloin tietokanta kasvaa prosessin edetessä ja osumia on jatkossa helpompi löytää. Yhdistelemällä erilaisia tekniikoita voidaan luoda haastavampia yhdistämisalgoritmeja kuten päätöspuurakenteita (decision tree) (Lennert, 2015).

Kaikki yhdistämisalgoritmit suorittavat merkkijonojen vertailua toisiinsa, joko merkki merkiltä tai tutkiakseen ovatko merkkijonot pääpiirteissään samat, jolloin pienet kirjoitusvirheet sallitaan. (Goldberg, 2008) Kaksi yleisimmin käytettyä merkkijonojen vertailua lyhyiden tekstien, kuten katujen ja kaupunkien nimien yhdistämistä varten, ovat Lennertin (2015) mukaan n-grammi ja Levenshteinin etäisyys. Recchia ja Louwerse (2013) vertailevat työssään 21 eri yhdistämisalgoritmia tai niiden variaatiota. He vertailevat miten menetelmät soveltuvat eri maiden paikannimien yhdistämiseen. Tutkimuksen mukaan vertailualgoritmien paremmuusjärjestys vaihtelee suuresti eri kieliryhmiin kuuluvien maiden välillä. Esimerkiksi hyvin Arabimaiden aineistoille toimiva menetelmä antaakin huonoja tuloksia kiinalaiselle ja ranskalaiselle aineistolle. Recchia ja Louwerse (2013) mukaan kannattaa, mikäli mahdollista, testata useita eri algoritmeja maa- tai kielikohtaiselle aineostolle ja jatkossa käyttää parhaiten toimivaa menetelmää.

Suomi ei ole mukana vertailussa Recchian ja Louwersen (2013) vertailussa ja siksi tässä työssä vertaillaan eri algoritmien toimivuutta pääasiassa suomenkielisellä aineistolla. Tut-

kittaviksi algoritmeiksi valikoituivat Levenshteinin etäisyys sekä n-grammin kaksi eri variaatiota ja niiden lisäksi Pisin yhteinen osa (longest common subsequence, LCS) ja ääntämiseen perustuva Soundex, jotka ovat myös yleisesti käytettyjä merkkijonojen vertailumenetelmiä.

2.2.1 Levenshteinin etäisyys

Levenshteinin etäisyys (Levenshtein distance tai edit distance) kuvaa editointietäisyyttä, eli kuinka monta kirjaimen poistoa, lisäystä tai vaihtoa on tehtävä, jotta sanoista tulee samat. Tällä tavalla pyritään siihen, että pienet kirjoitusvirheet eivät estä osuman löytymistä. (Järvelin ym. 2007; Lennert, 2015). Perusmuodossa jokaiselle kirjaimen muutokselle annetaan painoarvo 1, mutta erilaisille muutoksille voidaan myös antaa eri panoarvot niiden yleisyydestä riippuen (Järvelin ym. 2007). Levenshteinin etäisyys merkkijonojen a ja b (pituudet |a| ja |b|) välillä on $lev_{a,b}(|a|, |b|)$, jossa

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & jos\ min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & muutoin. \end{cases}$$

$1_{(a_i \neq b_j)}$ on 0 kun $a_i = b_j$ ja muutoin yhtä kuin 1. Min-osion ensimmäinen elementti vastaa kirjaimen poistoa, toinen lisäystä ja kolmas kirjaimen korvaamista. (Ranzijn, 2013).

Charif ym (2010) vertailevat työssään useita eri vertailualgoritmeja ja he pitävät niistä Levenshteinia luotettavimpana. Levenshtein ei kuitenkaan heidän mukaansa pysty käsittelemään lyhenteistä johtuvia virheitä, jonka vuoksi he yhdistävät sen ”vektoritekniikkaan”. Sillä vertailtavat nimet pilkotaan sanoiksi tai merkkijonoiksi, joita verrataan Levenshteinilla toisiinsa. Toinen vaihtoehto on verrata vain paria ensimmäistä kirjainta. Ranzijn (2013) mukaan Levenshteinin menetelmä soveltuu hyvin pääasiassa kirjoitusvirheiden korjaamiseen, mutta se on hidas pidempiä sanoja verratessa. On kuitenkin olemassa useita tekniikoita, jolla laskenta-aikaa voidaan lyhentää, kun merkkijonoa halutaan verrata kaikkiin vertailuaineiston sanoihin.

2.2.2 Pisin yhteinen osa

Pisin yhteinen osa (Longest common subsequense, LCS) menetelmällä lasketaan, kuinka paljon samoja merkkejä samassa järjestyksessä kahdessa vertailtavassa merkkijonossa esiintyy. Samanlaisuutta kuvaava vertailuarvo saadaan joko jakamalla pisin yhteinen merkkijono pidemmän merkkijonon pituudella, lyhyemmän merkkijonon pituudella tai vertailtavien merkkijonojen pituuksien keskiarvolla. Merkkien ei ole välttämättä oltava vierekkäisiä, mikä mahdollistaa merkkijonojen yhdistämisen, vaikka toiseen onkin eksynyt ylimääräinen merkki tai sieltä puuttuu jokin merkki. LCS osaa käsitellä useita virhetyyppejä, mutta se ei toimi vertailtavien merkkijonojen koostuessa useista sanoista, jotka ovat eri järjestyksessä. (Ranzijn, 2013).

Longest common substring toimii muuten samalla periaatteella, mutta siinä merkkien on oltava molemmissa merkkijonoissa sekä samassa järjestyksessä että vierekkäisiä. Molemmat menetelmät soveltuvat hyvin useiden eri virheiden korjaamiseen, riippuen siitä millä yllä mainituista tavoista samanlaisuusarvo lasketaan. (Ranzijn, 2013)

2.2.3 s-grammi ja n-grammi

S-grammi yhdistely (s-gram tai skip-gram matching) on merkkijonojen yhdistämisalgoritmi, jossa merkkijonot jaetaan tietyn mittaisiin osiin, eli s-grammeihin, joiden samanlaisuutta voidaan verrata. Menetelmä pohjautuu n-grammeihin, joita välillä kirjallisuudessa kutsutaan q-grammeiksi. N-grammissa merkkijonon osat koostuvat aina vierekkäisistä merkeistä, kun taas s-grammissa voidaan hypätä yli tietty määrä merkkejä s-grammeja muodostettaessa. N-grammia voidaan siis pitää s-grammin erityistapauksena. Yleensä merkkijonot jaetaan kahden tai kolmen merkin osiin, joita kutsutaan nimillä digrammi ja trigrammi. (Järvelin ym. 2007) N-grammin taustalla oleva ajatus on, että jos kaksi sanaa viittaavat samaan asiaan ja toisessa sanassa on kirjoitusvirheitä, on niissä todennäköisesti kuitenkin ainakin yksi n kirjaimen pituinen yhteinen merkkijono. N-grammin ei kuitenkaan pysty havaitsemaan kaikkia kirjoitusvirheitä ja toimii huonosti etenkin lyhyiden paljon virheitä sisältävien sanojen kohdalla. (Nikita, 2011) Esimerkkinä voidaan ajatella tapaus, jossa on ollut tarkoitus kirjoittaa Turku, mutta sormi on lipsahtanut ja siitä tulikin Tutku. Kun ”tutku” jaetaan trigrammeihin niistä tulee: tut, utk, tku. Mikään näistä trigrammeista ei esiinny sanan Turku trigrammeissa (tur, urk, rku), mistä johtuen menetelmä ei osaa yhdistää näitä merkkijonoja toisiinsa.

N-grammien samanlaisuutta voidaan kuvata erilaisilla kertoimilla. Yhteiset n-grammit voidaan jakaa merkkijonojen n-grammien kokonaismäärien keskiarvolla (Dice’s coefficient). Vaihtoehtoisesti yhteiset voidaan jakaa pitemmän merkkijono n-grammien lukumäärällä (Jaccard similarity) tai lyhyemmän merkkijonon n-grammien lukumäärällä (Overlap coefficient) (Ranzijn, 2013). Lennert (2015) puolestaan jakaa omassa tutkimuksessaan yhteiset n-grammit sanojen yhteispituudella. Grammeissa käytetään usein välilyöntejä täytteenä, jotta merkkijonojen alku ja loppupään kirjaimet saisivat samalaisen painoarvon vertailussa. Joissain tapauksissa saattaa kuitenkin olla parempi jättää täytteet käyttämättä, koska näin saadaan vähennettyä mahdollisten etu- ja jälkiliitteiden vaikutusta. Pelkästään sanan alun täyttäminen on havaittu hyödylliseksi suomenkielisten merkkijonojen kohdalla, johtuen suomen runsaasta jälkiliitteiden määrästä. (Järvelin ym., 2007; Ranzijn, 2013).

S-grammit voidaan jakaa luokkiin, sen perusteella monenko kirjaimen yli hypätään, niin että vain samaan luokkaan kuuluvia s-grammeja verrataan toisiinsa. Luokittelua kuvataan merkkikombinaatio indeksillä (character combination index, CCI) esimerkiksi CCI {{0},{1,2}} tarkoittaa merkkijonosta on tehty kaksi luokkaa. Yksi joka koostuu tavallisista n-grammeista, eli kahdesta peräkkäisestä merkistä, ja toinen jossa on kahdenlaisia s-grammeja, jotka on muodostettu hyppäämällä yhden tai kahden merkin yli. (Järvelin ym., 2007) Taulukko 1 esittää merkkijonon ”navetta” CCI merkintöjä ja s-grammeja ilman edellä mainittujen täytteiden käyttöä.

Taulukko 1. S-grammi luokat

Tyyppi	CCI	s-grammi luokat
s2,0	{0}	{na,av,ve,et,tt,ta}
s2,1	{1}	{nv,ae,vt,et,ta}
s2,2	{2}	{ne,at,vt,ea}

S-grammeja ei olla käytetty geokoodauksen yhdistämisalgoritmina, mutta sen sijaan niitä on käytetty muunlaiseen tietojen yhdistelyyn. Järvelin ym. (2007) tutkivat s-grammin käyttöä sukulaiskielissä esiintyvien sanojen yhdistelyssä. N-grammeja geokoodauksessa puolestaan ovat käyttäneet esimerkiksi Ranzijn (2013) sekä Lennert (2015). Lennertin (2015) tekemien

testien perusteella Levenshtein ja trigrammi antoivat samankaltaisia tuloksia, joskin trigrammi oli hieman parempi. Myös Ranzijn (2013) tutkimuksessa erityyppiset trigrammit antoivat parhaimmat tulokset.

2.2.4 Soundex

Merkkijonojen vertailuun voidaan myös käyttää sanojen ääntämiseen perustuvaa menetelmää, joista tunnetuin on Soundex-algoritmi. Sillä pyritään yhdistämään merkkijonoja, jotka kirjoitetaan eri tavalla mutta jotka kuitenkin kuulostavat samalta, muuntamalla ne ääntämistä kuvaavaksi koodiksi tiettyjen sääntöjen perusteella. (Goldberg, 2008) Koodin ensimmäinen merkki on muunnettavan sanan ensimmäinen kirjain ja sitä seuraa vähintään kolme numeroa. Soundex on aika herkkä mm. eroaville kirjoitusasuille, virheille ensimmäisen kirjaimen osalta ja konsonanttien paikan vaihdoille. (Zandbergen, 2008) Merkkijonojen muuttaminen koodiksi johtaa väistämättä tiedon menettämiseen, mikä saattaa johtaa virheellisiin niminen yhdistämiseen (taulukko 2).

Taulukko 2. Soundex esimerkki (Goldberg, 2008)

Original	Soundex
Running Ridge	R552 R320
Runs Ridge	R552 R320
Hawthorne Street	H650 S363
Heatherann Street	H650 S363

Soundexia olisikin paras käyttää mahdollisten tulosten suodatukseen vertailutietokannasta ja sen jälkeen käyttää jotain toista menetelmää lopullisen vastineen valitsemiseen. Suodatus voidaan tehdä esimerkiksi niin, että merkkijonot, joiden Soundexkoodi on tarpeeksi samankaltainen haettavan merkkijonon kanssa, otetaan mukaan vaihtoehdoiksi toisella menetelmällä tehtävään valintaan. (Ranzijn, 2013; Goldberg, 2008)

Alkujaan Soundex kehitettiin englanninkieltä varten, mutta sitä voidaan muokata myös muiden kielten käyttöön (Järvelin ym., 2007). Koska tässä työssä käytettävä aineisto on pääasiassa suomeksi, ei tässä testata Soundexin toimivuutta käytännössä. Geokoodausmenetelmää edelleen kehitettäessä saattaa kuitenkin olla hyvä harkita Soundexin ottamista osaksi sitä.

2.3 Luotettavuus

Pääasiallinen huolenaihe geokoodattaessa on tulosten luotettavuus. Erilaiset aineistot vaativat erilaisia menetelmiä. Isot kansainväliset aineistot sisältävät lukuisia erityyppisiä osoitteita, jolloin tarvitaan huomattava määrä sääntöjä sille, miten niiden osat luokitellaan, jotta niitä pystytään vertaamaan vertailuaineistoon. Pienet paikalliset aineistot puolestaan vaativat huomattavan paljon tarkempaa paikannustarkkuutta, sillä isossa mittakaavassa virheet sijainnissa korostuvat. (Zandbergen, 2008)

Tulosten luotettavuutta voidaan tutkia täydellisyyden, sijaintitarkkuuden ja toistettavuuden avulla. Täydellisyydellä tarkoitetaan osumaosuutta, eli kuinka suurelle osalle datasta onnistuttiin löytämään sijainti. Täydellisyys on helpoin tapa kuvailla geokoodaustuloksen laatua. Se on kuitenkin erittäin subjektiivinen mittari, sillä siihen vaikuttaa merkittävästi mitä kriteerejä osumalle asetetaan. Laskemalla osuman kriteerejä saadaan nostettua osumien määrää, mutta samalla voidaan saada enemmän virheellisisä tuloksia, jolloin luotettavuus heikkenee. Sijaintitarkkuus kuvaa geokoodatun pisteen ja osoitteen todellisen sijainnin euklidista

etäisyyttä. Toistettavuus puolestaan merkitsee prosessin herkkyyttä muutoksille esimerkiksi algoritmeissa ja geokoodaajan taidoissa. (Zandbergen, 2008; Zandbergen, 2009)

Geokoodauksen tuloksen luotettavuuteen vaikuttavat useat tekijät, kuten osoitteiden maantieteelliset alueet, vertailuaineiston laatu sekä käytettävät algoritmit (Roongpiboonsopit & Karimi, 2010). Yhdistämisalgoritmilla on keskeinen vaikutus tuloksen laatuun, sillä se määrittää lopputuloksen johtamiseen käytettävän kohteen (Goldberg, 2008). Osoitteiden geokoodauksessa myös rakennuksen käyttötarkoituksella on vaikutusta geokoodauksen tuloksiin, sillä asuinrakennuksilla on Zandbergenin (2008) tutkimuksessa selvästi parempi osumatarkkuus kuin liikerakennuksilla. Cayo ja Talbotin (2003) tekemän tutkimuksen mukaan katuaineistojen (street reference files) avulla geokoodattujen sijaintien ja osoitteiden todellisten sijaitien välillä on huomattavia eroja. Yksi syy siihen voi heidän mukaansa olla, että katuverkkoaineisto sisältää täydellisempiä osoitetietoja tiheämmin rakennetuilla alueilla.

Goldberg ym. (2007) esittelee useita tulokseen vaikuttavia näkökulmia, jotka tulee ottaa huomioon geokoodausmenetelmää valittaessa. Vertailuaineiston tarkkuus vaikuttaa huomattavasti tuloksen tarkkuuteen. Tässä on otettava huomioon paitsi sijaintitarkkuus, myös aineiston ajantasaisuus sekä aineiston alkuperäinen käyttötarkoitus. Lisäksi geokoodauksen tuloksen pinta-yksikön (areal unit) valinta, eli onko tuloksena yksittäinen piste vai postinumeroalue, vaikuttaa tulokseen. Sekä syötteessä että vertailuaineistossa voi esiintyä kirjoitusvirheitä, epäjohdonmukaisuuksia osoitteen etuliitteessä tai lyhenteissä sekä virheitä postinumerotiedoissa (Cayo ja Talbot, 2003). Cayo ja Talbot (2003) havaitsivat tutkimuksessaan syyn yli viiden kilometrin virheisiin sijainnissa olevan usein epätarkkuus vertailuaineiston postinumeroalueissa. Tästä syystä osoite saatettiin paikantaa viereiseen postinumeroalueeseen kilometrien päähän. Ajantasaisuuden osalta on otettava huomioon, että haettavan osoitteen ja vertailuaineiston on oltava samalta ajalta, sillä esimerkiksi kunnanrajojen muuttuminen saattaa vaikuttaa siten, ettei vanhalle osoitteelle löydy vastinetta uuteen vertailuaineistoon verrattaessa.

Useiden tutkimusten mukaan maaseudulla on huomattavasti heikompi osumatarkkuus kuin kaupunkialueilla sijaitsevilla osoitteilla (Roongpiboonsopit & Karimi, 2010) ja siksi tulosaineiston laatu saattaa vaihdella eri alueilla. Yksi syy tähän on esimerkiksi Yhdysvalloissa käytettävät postilokero-osoitteet, jotka saattavat sijaita kaukana rakennuksesta, jonka koordinaatteja etsitään (Zandbergen, 2009). Toinen syy tähän on, että lineaarinen interpolointi, jota kuvataan tarkemmin luvussa 2.4.2, antaa tarkempia tuloksia mitä lyhyempiä kadut ovat ja maaseudulla matka kahden risteyksen välillä on pääsääntöisesti huomattavasti pidempi. (Goldberg, 2008) Erilaiset algoritmit tekevät erilaisia olettamuksia ja näiden sopivuutta omaan tarpeeseen on syytä tarkastella. Esimerkiksi lineaarisen interpolointiin perustuva geokoodaus olettaa, että kaikki osoitteet tietyllä numerovälillä ovat olemassa. Geokoodausmenetelmää valittaessa on myös otettava huomioon, hyväksytäänkö tulokseen epävarmoja tuloksia, eli käytetäänkö todennäköisyyspohjaista vai determinististä yhdistämisalgoritmia. (Goldberg ym., 2007)

Tulosten luotettavuus vaikuttaa geokoodatun aineiston perusteella tehtävien paikkatietoanalyysien luotettavuuteen. Tulosten osumaosuus vaikuttaa analyyseissä käytettävään aineistoon ja sen myötä tulokseen. Mikäli tietyllä alueella ei onnistuta paikantamaan riittävää määrää osoitteita kärsii alueen otanta. Geokoodauksesta aiheutunut systemaattinen virhe saattaa johtaa tulokseen, jossa esimerkiksi terveydentilan ja asuinpaikan välillä havaitaan yhteys, vaikka sitä ei oikeasti ole olemassa. (Cayo ja Talbot, 2003)

Tulosten sijaintitarkkuutta voidaan mitata eri tavoin, joista osa on halvempia ja toiset erittäin kalliita. Luotettavin tapa on mennä GPS-laitteen kanssa paikan päälle ja mitata jokaisen osoitteen koordinaatit. Tämä on kuitenkin mahdollista vain jos tarkistettava datamäärä on pieni ja sijoittuu maantieteellisesti rajatulle alueelle, sillä muussa tapauksessa vaadittava työmäärä on valtava. Useimmiten on siis turvauduttava muihin keinoihin. Yksinkertaisinta on verrata uuden geokoodauksen tuottamaa tulosta aikaisemman hyväksi havaitun geokoodauksen tulokseen. Tätä menetelmää käytetään etenkin uusia geokoodaussovelluksia kehitettäessä. Toinen vaihtoehto on käyttää ilmakuvia, joilta pystytään näkemään kohde, jota koordinaattien tulisi vastata ja mittaamaan, kuinka kaukana se on geokoodauksen koordinaateista. On kuitenkin tärkeää pitää mielessä, että ilmakuvien georeferoinnissa saattaa olla virheitä, mutta niitä pidetään kuitenkin luotettavampina kuin geokoodauksen aikana interpoloituja kohteita. (Goldberg, 2008)

2.4 Osoitteiden geokoodaus

2.4.1 Sovellusalueet

Geokoodattua dataa voidaan hyödyntää analyyseihin useilla eri tieteenaloilla. Lääketieteessä voidaan yrittää havaita tietyn taudin alueellisia keskittymiä tai selvittää mihin uusi terveyskeskus tulisi sijoittaa. Yhdistettynä navigointijärjestelmään tarkasti geokoodattu data voi auttaa hätäkeskusta löytämään onnettomuuspaikan, mikäli annettu osoite on joltain osin puutteellinen. Kaupan alalla geokoodattua asiakasdataa voidaan yhdistää demograafiseen aineistoon, jotta voidaan selvittää asiakaspohjaa uudelle liikkeelle. (Christen ym., 2004) Geokoodaus onkin saanut paljon huomiota tieteellisissä tutkimuksissa jo usean vuoden ajan, ei ainoastaan geoinformatiikan alalla vaan esimerkiksi historiassa, terveystieteessä ja kriminologiassa. Yksi syy tähän on se suuri määrä paikkatietoa, joka osoitteista voidaan saada ja jota voidaan hyödyntää erilaisissa paikkatietoanalyyseissä. (Goldberg, 2011)

Christen ym. (2004) mukaan on olemassa kaksi pääasiallista tapausta, joissa geokoodausta käytetään. Ensimmäisessä käyttäjällä on aineisto, jonka hän haluaa geokoodata automaattisesti. Järjestelmän tulisi tällöin löytää mahdollisimman hyvä sijainti jokaiselle tietueelle, sekä tieto siitä kuinka tarkka kyseinen sijainti on, ilman että ihmisen tarvitsee tehdä manuaalista työtä. Toisessa tapauksessa henkilö haluaa geokoodata yhden ainoan osoitteen mahdollisimman tarkasti ja nopeasti. Tuloksen saaminen saa kestää korkeintaan sekunteja ja siihen käytetään yleensä internetistä löytyvää palvelua. Järjestelmän tulisi palauttaa tarkka vastaus, tai mikäli sellaista ei löydy, lista mahdollisista sijainneista ja tiedot kunkin sijainnin todennäköisyydestä. (Christen ym., 2004). Näistä tapauksista ensimmäinen on se mihin tässä työssä keskitytään.

Osoitteiden geokoodaukseen on olemassa useita sovelluksia, sekä kaupallisia, kuten ArcMap, että web-pohjaisia, kuten Google Maps, Yahoo! ja Bing. Ilmaiset web-sovellukset kuitenkin rajoittavat tulosten käyttöä siten, että ainoastaan tulosten tarkastelu kartalla on mahdollista ja analyyttisen tarkastelun mahdollisuus on suljettu pois. Syy tähän ei ole itse palvelussa, vaan sen taustalla olevassa datassa, jonka yksityiset tahot ovat tuottaneet. Yhdysvalloissa on olemassa ilmaisia vaihtoehtoja, jotka perustuvat Census Bureau Tiger dataan, mutta Euroopassa vastaavan kaltaiseen dataan käsiksi pääseminen on ollut Lennertin (2015) mukaan viime vuosiin saakka haastavaa. OpenStreetMap joka perustuu yleisön osallistamiseen datan keräämiseen (crowd-sourcing) on kuitenkin viime vuosina helpottanut tätä ongelmaa.

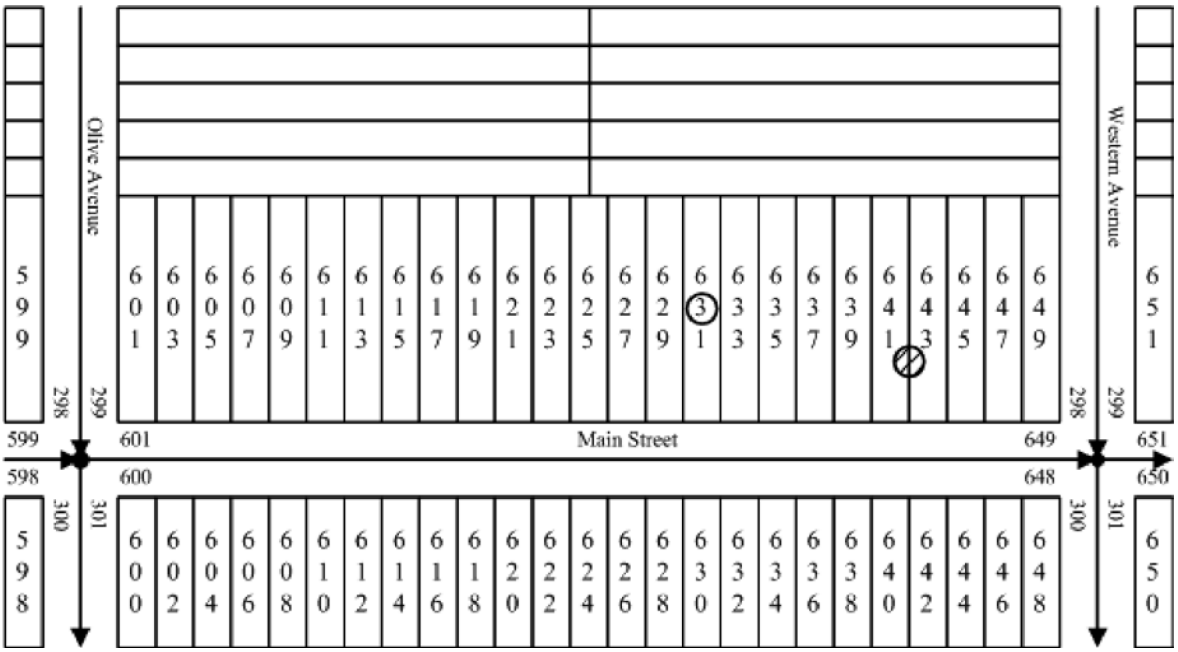
Avoimen datan määrä on lisääntynyt huomattavasti vuonna 2007 voimaan tulleen EU:n Inspire-direktiivin myötä. Siinä määrätään joukko aineistoja, joita on tarjottava joko lataus tai katselupalveluna ja osoitteet ovat yksi tärkeimmistä aineistoista. (Lennert, 2015)

2.4.2 Katuverkkomenetelmä

Katuverkkoon perustuva geokoodaus on yleisimmin käytetty osoitteiden geokoodausmenetelmä (Zandbergen 2008; Charif ym. 2010). Siinä katuverkko on jaettu segmentteihin, joilla on tiedossa kadun nimi sekä molemmissa päissä olevat talonnumerot, pääsääntöisesti erikseen kummallekin puolelle tietä. Geokoodattavaa osoitetta verrataan ensin kadun nimeen ja sen jälkeen talon numeroon. Kun oikea katusegmentti on löydetty, käytetään lineaarista interpolointia, jolla piste sijoitetaan katusegmentin varrelle. Piste voidaan sijoittaa tienviivasta hiukan sivuun, jotta on mahdollista erottaa, kummalla puolella katua piste sijaitsee. Katuverkkogeokoodaus ei tarkista onko haettu osoitenumero oikeasti olemassa ja siksi interpolointi johtaa virheelliseen tulokseen esimerkiksi tilanteessa jossa haettu osoitenumero osuu viivasegmentin numeroiden alueelle, mutta ei ole todellinen osoite. Menetelmän osumatarkkuus vaihtelee noin 70-95 % välillä, mutta luku saattaa siis sisältää virheellisisä osumia (Zandbergen, 2009). Zandbergenin (2008) mukaan melkein kaikki sekä online geokoodausjärjestelmät että kaupalliset geokoodaus ohjelmat käyttävät katuverkkomenetelmää.

Katuverkkogeokoodaukseen käytettävä vertailuaineisto voi koostua joko suorista janoista tai murtoviivoista. Murtoviivat tuottavat yleensä paremman tuloksen, sillä ne kuvaavat kadun muotoa paremmin kuin yksittäinen suora viiva. Useimmiten katuverkko esitetään graafina, joka koostuu solmupisteistä sekä niiden välisistä särmistä. (Goldberg, 2008)

Lineaarista interpolointia varten tehdään muutamia oletuksia, jotka saattavat kuitenkin vääristää tulosta. Ensinnäkin oletetaan, että kaikki katusegmentin alku- ja loppupään numeroiden välissä olevat osoitenumerot ovat todellisia käytössä olevia osoitteita. Toinen oletus on, että kiinteistöt ovat saman kokoisia ja ulottuvat siten yhtä pitkälle matkalle kadun varrella. Kolmanneksi oletetaan asuttujen kiinteistöjen ulottuvan kadun päästä päähän, niin ettei väliin jää tyhjiä alueita. Nämä oletukset ja niiden vaikutus on esitetty kuvassa 2. Ensimmäisen oletuksen vaikutusta voidaan vähentää lisäämällä vertailuaineistoon tieto katusegmentin varrella oikeasti olevien kiinteistöjen lukumäärästä, mutta tämä ei kuitenkaan vielä poista muita olettamia. Koska todellinen tilanne harvoin on oletusten mukainen johtavat ne virheeseen, jonka suuruus riippuu segmentin pituudesta. (Goldberg ym, 2007; Goldberg, 2008)



Kuva 2. Katuverkkogeokoodauksen oletusten aiheuttama virhe. Geokoodauksen tulos esitetty ympyrällä ja todellinen sijainti pisteellä. (Goldberg ym, 2007)

Edellä mainittujen oletusten lisäksi tulokseen vaikuttaa kadunkulman kiinteistö, sillä pääsääntöisesti ei ole tiedossa minkä tien osoitteisiin se kuuluu (Goldberg, 2008). Kulmaongelmaa voidaan pyrkiä korjaamaan päätysiirtymällä (end offset), jolloin kadun varren numerot puristetaan kasaan kadun pituussuunnassa, eli interpolointia ei tehdä koko kadun mitalle. Tällöin reunimmaisten osoitteiden sijoittuminen risteysalueelle vähenee ja sijaintitarkkuus paranee. Päätysiirtymänä voidaan käyttää joko tiettyä metrimäärää tai prosenttiosuutta katusegmentin pituudesta. Koska katuverkko pääsääntöisesti kuvaa tien keskiviivaa, tehdään osoitepisteelle lisäksi katuun nähden kohtisuora siirtymä, jonka suuruus on yleensä 10-15 m, vaikkakin asuinpaikka todellisuudessa saattaa sijaita huomattavasti kauempana kadusta. (Zandbergen, 2009) Cayon ja Talbotin (2003) mukaan optimaaliset siirtymät olisivat 15 metriä kadusta ja 50 m risteyksestä. Mainituilla siirtymillä on kuitenkin havaittu olevan kohtalaisen vähän vaikutusta sijaintitarkkuuteen, sillä ne vähentävät virhettä keskimäärin vain muutamia metrejä.

Katuverkkomenetelmän sijaintitarkkuuteen vaikuttavat, paitsi oletuksista johtuvat virheet, myös virheet katusegmenteissä. Virhe vertailuaineiston osoitetiedoissa ja sijainnissa tai haettavassa osoitteessa saattaa johtaa osoitteen yhdistämiseen kokonaan väärään segmenttiin. Tällöin virhe voi olla useita kilometrejä, eikä pienillä korjauksilla tuolloin ole merkitystä. Mikäli osoite on onnistuttu paikantamaan oikeaan kohtaan katua parantavat siirtymät yleensä tuloksia hieman, mutta jos osoite on alun perin paikannettu väärään kohtaan katua saattaa siirtymä vain huonontaa tulosta. (Zandbergen, 2009) Yksi syy siihen miksi harvaanasutuilla alueilla on havaittu keskustoja huonompaa sijaintitarkkuutta, on että etäisyydet risteyksien välillä ovat suurempia, jolloin katusegmenttien pituus kasvaa. Mitä pidempää segmenttiä interpoloinnissa käytetään, sitä huonompi sen tarkkuus on. (Cayo ja Talbot, 2003)

2.4.3 Aluemenetelmä

Aluerajamenetelmässä (parcel boundaries) osoitteet sovitetaan yhteen yksittäisten maa-alueita kuvaavien polygonien, kuten kiinteistöjen tai postinumeroalueiden, tai alueiden keskipisteiden kanssa. Tämä on hyödyllistä etenkin alueilla, joissa rakennus sijaitsee kauempana tien keskiviivasta. Koska osuma saadaan vain kun täsmäävä alue löytyy, saadaan aluegeokoodauksella pääsääntöisesti vähemmän osumia, kuin katuverkkomenetelmällä, yleensä noin 40-75 %. Toinen syy harvempiin osumiin on, että alue voidaan yhdistää useampiin osoitteisiin, kuten esimerkiksi asunto-osakeyhtiössä, mutta jokaisen asunnon osoitetta ei välttämättä ole vertailuaineistossa. Pienemmästä osumaosuudesta huolimatta aluegeokoodausta pidetään sijaintitarkkuudeltaan parempana kuin katuverkkogeokoodausta. (Zandbergen, 2008) Yksi syy tähän voi olla aineiston päivityksessä. Koska kiinteistönrajat vaikuttavat verotukseen pidetään ne ajan tasalla, kun taas tieverkkoa kuvaavat aineistot päivitetään harvemmin. Toisaalta aluedatan osoitteita keräävät monet paikalliset tahot, jonka vuoksi niitä ei välttämättä ole tallennettu samojen periaatteiden mukaisesti. (Cayo ja Talbot, 2003) Aluegeokoodauksen käyttö on lisääntynyt, mutta sitä käytetään lähinnä maantieteellisesti rajallisilla alueilla, johtuen laajojen aineistojen puutteesta ainakin Yhdysvalloissa. Aluegeokoodaus on haastavampaa kuin katuverkon käyttäminen, johtuen muun muassa siitä, ettei aluedataa ole varsinaisesti suunniteltu geokoodausta varten. (Zandbergen, 2008)

Polygonidata saattaa vaihdella hyvin tarkasta vertailuaineistosta kovinkin epätarkkaan. Aineisto joka kuvaa rakennuksen alaa on yleensä hyvin tarkkaa, mutta vaikka niiden perusteella tehtyjä karttoja on paljon, niin taustalla oleva data on harvoin saatavissa geokoodauskäyttöön. Kiinteistöjen rajat puolestaan ovat paremmin saatavilla ja paikoissa joissa ne on mitattu GPS:llä ovat nekin erittäin tarkkoja. Polygoneja jotka kuvaavat kaupunkeja ei huonon tarkkuuden vuoksi voida välttämättä käyttää geokoodauksen tuloksina, mutta niitä voidaan käyttää rajaamaan alue esimerkiksi etsittäessä segmenttiä katuverkkoaineistosta. (Goldberg, 2008)

Jos geokoodauksesta halutaan saada tulokseksi yksi koordinaatti eikä koko polygonia, on yhdistämisalgoritmin valitsemalle alueelle suoritettava interpolointi samaan tapaan kuin katusegmentille. Yksinkertaisin menetelmä on laskea alueen sijaintia rajaava suorakaide (bounding box) ja sijoittaa piste sen keskipisteeseen. Tämä toimii alueen ollessa suorakaiteen muotoinen, mutta jos muoto on kovin epäsäännöllinen, saattaa käydä niin että keskipiste sijaitseekin alueen ulkopuolella. Hieman haastavampi vaihtoehto on laskea alueen maantieteellinen keskipiste. Tällöin varmistetaan, että keskipiste on alueen sisällä, mutta samalla menetelmä on laskennallisesti vaativampi kuin edellä mainittu. Tämä menetelmä toimii kohtalaisen hyvin tiiviisti rakennetuilla alueilla, jossa rakennus usein on melko keskellä kiinteistöä, mutta mitä isompi alue on, sitä kauempana sen keskipiste on rakennuksesta, mikä voidaan huomata esimerkiksi maatilojen kohdalla. Kun vertailuaineistona käytetään isoja alueita, Goldberg suosittelee käyttämään painotetun keskipisteen laskevaa menetelmää, esimerkiksi postinumeroalueiden toimiessa vertailuaineistona voidaan niihin yhdistää väestöntiheyttä kuvaava tarkempi aineisto rasterimuodossa. Tällöin alueen painotettu keskipiste voidaan laskea väestön mukaan, jolloin piste on lähimpänä paikkaa jossa useimmat ihmiset oikeasti asuvat. (Goldberg, 2008; Goldberg, 2013)

2.4.4 Osoitepistemenetelmä

Osoitepistemenetelmässä (address point) käytettävä data voidaan luoda esimerkiksi käyttämällä alue-datan keskipisteitä tai sijoittamalla piste alueen päärakennuksen kohdalle sekä täydentämällä sitä yksittäisten asuntojen osoitteilla. Kun aluedatassa on vain yksi alue kuvaamassa kokonaista rivitaloa, voidaan osoitepistemenetelmällä kuvata jokaista asuntoa erikseen. Pisteaineistoa voidaan täydentää ilmakuvilla, jolloin pisteet saadaan sijoitettua rakennuksen keskelle alueen keskipisteen sijaan. Samoin kuten aluemenetelmässä tässäkin osumat vaativat kyseisen osoitepisteen löytymistä. (Zandbergen, 2008) Esimerkiksi Australiassa ja Isossa-Britanniassa on kehitetty osoitepistetietokantoja, mutta niitä on toistaiseksi tutkittu aika vähän ja menetelmä on vähäisessä käytössä. Suomessa Maanmittauslaitoksen maastotietokantaan sisältyy osoitteiden kyselypalvelu (WFS), jonka kautta on mahdollista saada osoitepisteet, mutta sen käyttö on maksullista ja edellyttää sopimusta (Maanmittauslaitos, 2017). Zandbergen (2008) pitää menetelmää lupaavana johtuen menetelmän hyvästä sijaintitarkkuudesta sekä siitä, että osumaosuus on samaa luokkaa katuverkkomenetelmän kanssa. Lisäksi virheellisten osumien todennäköisyys on pienempi kuin katuverkkomenetelmässä. Osumien iso osuus saattaa kuitenkin selittyä osittain sillä, että osoitepisteaineistot yleensä tehdään geokoodausta varten.

2.4.5 Yhdistelmä

Nykypäivänä useimmat geokoodaussovellukset käyttävät useita eri vertailuaineistoja, jolloin jokainen taso sisältää yhdenlaisia kohteita tietyssä mittakaavassa. Tasoja voivat olla esimerkiksi rakennusten keskipisteet, kiinteistöt, katusegmentit sekä postinumeroalueiden, läänien ja maiden keskipisteet. Vaihtoehtoisesti saman tyyppisiä kohteita voidaan esittää useilla eri tasoilla jaoteltuina aineiston tuottajan, tuotantotavan tai aineiston esittämän ajankohdan mukaan. Tällä tavoin pystytään varmistamaan, että geokoodauksen osumaosuus on korkea, sillä jokaista osoitetta kohden voidaan palauttaa ainakin jonkun tasoinen kohde. Kohteiden sijaintitarkkuus kuitenkin heikkenee nopeasti vertailutason muuttuessa, jolloin käyttäjän saama hyöty kohteesta vähenee. (Goldberg & Cockburn, 2010) Mitä hyötyä on esimerkiksi tietää maan keskipiste, kun oikeasti haluttiin osoitteen sijainti.

Yksi esimerkki tällaisesta geokoodaukseen käytettävästä tietokannasta on Australialainen G-NAF, johon on kerätty yhteen lukuisien eri tahojen osoitetiedot. Sitä käyttävä geokoodausjärjestelmä palauttaa ensisijaisesti tarkan osuman. Mikäli oikeaa osoitenumeroa ei kuitenkaan löydy, se palauttaa tietä kuvaavat koordinaatit, jotka löytyvät toisesta tiedostosta. Jos oikeaa tietäkään ei löydy, haetaan paikkakuntatiedostosta kaupunkia tai lähiötä kuvaavat koordinaatit. (Christen ym., 2004) Useiden eri vertailuaineistojen käyttäminen saattaa johtaa tilanteeseen, jossa järjestelmällä on tiedossa useita eri kohteita joita se voisi palauttaa ja sen on niistä pystyttävä valitsemaan kohde, joka parhaiten vastaa käyttäjän tarvetta. (Goldberg & Cockburn, 2010)

2.5 Twittergeokoodaus

2.5.1 Twiittien paikannuksen perusteet

Sosiaalinen media on jokapäiväinen osa monen ihmisen elämää ja siksi sitä on alettu yhä enemmän hyödyntämään tutkimuksissa. Osoitteiden lisäksi viime vuosina on alettu geokoodata esimerkiksi Twitteristä ja Facebookista saatavia tietoja, joita voidaan hyödyntää tiedonhaussa monella alalla. Esimerkiksi Japanissa sattuneen maanjäristyksen jälkeen ihmiset käyttivät Twitteriä avun pyytämiseen puhelinlinjojen ollessa poikki (Zhang & Gelernter, 2014). Twiittien paikantaminen auttaa avun saannissa kriisitilanteissa, mutta sitä voidaan

käyttää myös mainosten kohdentamiseen (Han ym., 2014) tai vaalitulosten ennustamiseen (Zhang & Gelernter, 2014).

Vuonna 2016 twitterissä lähetettiin päivittäin noin 500 miljoonaa twiittiä (Sidkar ja Gambäck, 2016). Twiitit ovat korkeintaan 140 merkin pituisia tekstejä, jotka voivat koskea mitä tahansa ja niihin on mahdollista liittää mobiililaitteen GPS:n antamat koordinaatit twiittamishetkellä. On olemassa lukuisia sovelluksia, jotka visualisoivat twiittejä kartalla niiden koordinaattien perusteella, mutta vain noin 1 % kaikista twiiteistä sisältää koordinaatit ja siksi on alettu kehittää muita tapoja niiden paikantamiseen. Twiitin paikantamista sen sisällön perusteella on tällä vuosikymmenellä käsitelty useissa tutkimuksissa. Paikantamista varten twiitistä pyritään erottelemaan tekstissä tai twiitin metatiedoissa esiintyviä paikannimiä geokoodattaviksi. (Zhang & Gelernter, 2014; Han ym., 2014) Paikannimien erottamista aineistosta kutsutaan nimellä geoparsing ja se on osa tiedon eristämisprosessia (information extraction), jossa jäsentelemätöntä tietoa pyritään jäsentelemään. (Freire ym., 2011)

Geokoodattaessa paikannimiä, yksi ongelma osoitteisiin verrattuna on, että samalla paikannimellä voidaan viitata useaan sijaintiin eri puolilla maailmaa. Oikean vaihtoehdon valitsemiseen on hyödynnettävä muuta twiitistä saatavaa tietoa, kuten muita paikannimiä tai tiettyyn alueeseen viittaavia sanoja. Paikannimien lisäksi geokoodamiseen voidaan käyttää twiitin metatietoja, joita voivat koordinaattien lisäksi olla käyttäjän twittertilin avaamisen yhteydessä syöttämä kotipaikka tai muu sijaintiin viittaava tieto sekä aikavyöhyke, jolta twiittaus on tehty. (Zhang & Gelernter, 2014) Twiittien metatietoihin kuuluu lisäksi twitterin automaattisesti tunnistama tekstin kieli. Twiiteissä käytetään usein lyhenteitä, puhekieltä, erikoismerkkejä sekä yhdistetään useita sanoja yhdeksi. Lisäksi twiitit julkaistaan usein spontaanisti mobiililaitteella, jolloin oikeinkirjoituksentarkistus on heikompaa ja kirjoitusvirheitä sattuu herkästi. Myöskään isojen kirjainten käyttö ei ole läheskään niin säännönmukaista kuin virallisemmissa teksteissä. Kaikki tämä johtaa siihen, että paikannimien tunnistaminen ja yhdistäminen geokoodauksessa käytettävään vertailuaineistoon on hankalaa verrattuna osoitteiden geokoodaamiseen. (Han ym., 2014; Sidkar ja Gambäck, 2016)

Useimmat twitterin geokoodausta koskevat tutkimukset käsittelevät käyttäjän kotipaikan selvittämistä käyttämällä useita twiittejä samalta käyttäjältä. Voidaan olettaa, että jos käyttäjä toistuvasti mainitsee twiiteissään tietyn paikan, niin hän todennäköisesti twiittaa kyseiseltä alueelta. Tähän olettamukseen perustuvat useat sekä sanojen vertailuun, sääntöihin että koneoppimiseen (machine learning) perustuvat menetelmät (Han ym., 2014), joita kuvataan tarkemmin seuraavissa luvuissa. Twitterin käyttäjän kotipaikasta ollaan kiinnostuneita esimerkiksi tutkimuksissa, joissa halutaan vertailla ihmisten mielipiteitä ja uskomuksia. Yksittäisten twiittien paikantaminen puolestaan on tavoitteena esimerkiksi katastrofien pelastustöiden yhteydessä. (Alex ym., 2016)

Tulosten arviointiin ja menetelmien kehittämiseen on pääsääntöisesti käytetty valmiiksi georeferoituja twiittejä, eli niitä joihin käyttäjä on julkaissut koordinaatit puhelimestaan. Han ym. (2014) mukaan ei ole kuitenkaan täysin selvää miten hyvin georeferoidut twiiti vastaavat muita twiittejä, johtuen mm. siitä että georeferoidut twiitit on kirjoitettu puhelimella, kun taas muut on voitu kirjoittaa millä tahansa laitteella, mukaan lukien pöytäkoneella. Näin ollen on epävarmaa kuinka hyvin menetelmät, joita on testattu georeferoiduilla twiiteillä sopivat käytettäviksi muille twiiteille. (Han ym., 2014)

Han ym. (2014) hyödyntävät omassa tutkimuksessaan twiitin tekstin lisäksi metatietoja, jotka kertovat käyttäjän kotipaikan, aikavyöhykkeen, kuvauksen ja oikean nimen. Näistä ainoastaan teksti ja käyttäjän nimi ovat olemassa jokaisen twiitin kohdalla. Vaikka käyttäjän nimi ei suoranaisesti viittaa mihinkään paikkaan saattaa se kuitenkin joissain tapauksissa antaa vinkkejä, esimerkiksi Petrov saattaa olla Venäjällä tavallisempi nimi kuin muualla maailmassa ja Hasegawa puolestaan Japanissa. Kotipaikkaan ja kuvaukseen käyttäjä voi kirjoittaa mitä itse haluaa ja siksi niiden antama informaatio on hyvin vaihtelevaa. Osa ilmoittaa kotipaikkansa muodossa ”kaupunki, maa”, mutta hyvin useat kirjoittavat jotain aivan muuta, kuten ”paras paikka maailmassa”, tai käyttävät paikoista lyhenteitä ja lempinimiä. Käyttäjän on myös mahdollista muokata näitä tietoja. Han ym. (2014) tutkimuksesta käy ilmi, että noin 18 % käyttäjistä muutti kuvaustaan twitterissä noin viiden kuukauden tarkastelujakson aikana, kun taas esimerkiksi kotipaikkaa muutettiin alle 8 % twiiteistä. On hyvin mahdollista, etteivät käyttäjät päivitä kotisijaintia muuttaessaan, jonka seurauksena tieto saattaa olla vanhentunutta. Alex ym. (2016) mukaan käyttäjän kotipaikan selvittämisessä twiitin kotipaikkakentän tiedot antavat kuitenkin huomattavasti tarkemman kuvan henkilön asuinpaikasta, kuin mitä twiittien teksteistä kerätyt paikkailmaisut. Tämä johtuu siitä, että heidän tutkimusaineistossaan peräti 66 % satunnaisesti tarkistetuista 10 000 englanninkielisestä käyttäjäprofiilista sisälsi todellisen paikan, yleensä kaupungin, joka pystyttiin paikantamaan. Yksittäiset twiitit joista käy ilmi käyttäjän senhetkinen sijainti ovat puolestaan paljon harvemmassa.

Tekstin yhdistämistä sijaintiin voidaan hyödyntää twiittien lisäksi internetsivuilla, blogeissa, uutisartikkeleissa ja matkakertomuksissa (Leidner & Lieberman, 2011). Muutamia poikkeuksia lukuun ottamatta useimmat tekstin paikantamismenetelmät on tehty englanninkielisiä tekstejä varten, joko poistamalla muun kieliset twiitit aineistosta, tai valitsemalla twiittit pääasiassa englanninkielisiltä alueilta. Geokoodausmenetelmien kohdalla ei ole tutkittu miten hyvin ne toimivat muilla kielillä tai kun tekstit ovat useilla eri kielillä. Twitter on kuitenkin monikielinen media ja jotkut kielet saattavat antaa selkeitä viitteitä siitä, mistä päin maailmaa twiittaus on tehty. Jos käyttäjä esimerkiksi twiittaa jatkuvasti japaniksi on hyvin todennäköistä, että hän asuu Japanissa. (Han ym., 2014) Han ym. (2014) havaintojen mukaan englanniksi twiittaavat henkilöt lisäävät twiittiin koordinaatit hieman useammin kuin esimerkiksi japanilaiset, korealaiset ja saksalaiset. Eron voidaan spekuloida johtuvan edellä mainittujen kansalaisuuksien isommasta halusta varjella yksityisyyttään.

Muutkin sanat kuin paikannimet voivat viitata paikaan. Sanasto vaihtelee eri paikkakunnilla ja twitterin käyttäjä Lontoossa käyttää todennäköisemmin sanoja kuten ”tube” ja ”Piccadilly” kuin käyttäjä New Yorkissa tai Pekingissä. Tämä ei kuitenkaan tarkoita, etteikö näitä sanoja voisi käyttää missä tahansa muuallakin, vaan niiden esiintymistodennäköisyys on suurempi juuri Lontoossa. (Han ym., 2014) Myös twiitti joka koskee vuoden 2012 kesäolympialaisia voidaan yhdistää Lontooseen. Jotta tiettyyn paikkaan liittyviä sanoja voidaan käyttää twiittien paikantamisessa, on järjestelmän pystyttävä yhdistämään ne paikkaan. (Zhang & Gelernter, 2014) Paikkakuntakohtaiset sanat eivät ole yksiselitteisiä. Jos twiitissä mainitaan HIFK ei voida olettaa, että se on lähetetty Helsingistä, sillä joukkueella on saattanut olla vieraspeli jollain muulla paikkakunnalla.

Twiittien paikannuksessa sijaintitarkkuus on aivan eri luokkaa kuin osoitteiden kohdalla. Kun osoitteiden kohdalla koordinaatit pyritään saamaan oikean talon kohdalle, on twiittien osalta tärkeintä osua, tapauksesta riippuen, joko oikeaan kaupunkiin tai maahan, sillä niitä twiiteissä useimmiten mainitaan. Tarkempaankin on toki mahdollista päästä, esimerkiksi jos

twiitissä mainitaan eduskuntatalo ja vertailuaineistosta löytyy kyseiselle rakennukselle koordinaatit.

Dredze ym. (2016) tutkivat ajan vaikutusta yksittäisten twiittien geokoodaukseen. Yksittäisiä twiittejä tutkittaessa ajalla on merkitystä, sillä niiden julkaisupaikka vaihtelee kellonajan mukaan. Aamulla niitä lähetetään usein kotoa ja päivällä vuorostaan töistä. He havaitsivat, että twiittien geokoodauksen sijaintitarkkuus vaihtelee syklisesti ajan mukaan. Isot opetusaineistot parantavat tarkkuutta, mutta niiden hyödyt katoavat kun mallia käytetään uudempien twiittien paikantamiseen. Yksi tapa ylläpitää tarkkuutta on jatkuvasti ladata geokoodattuja twiittejä ja käyttää niitä opetusaineistona. (Dredze ym., 2016)

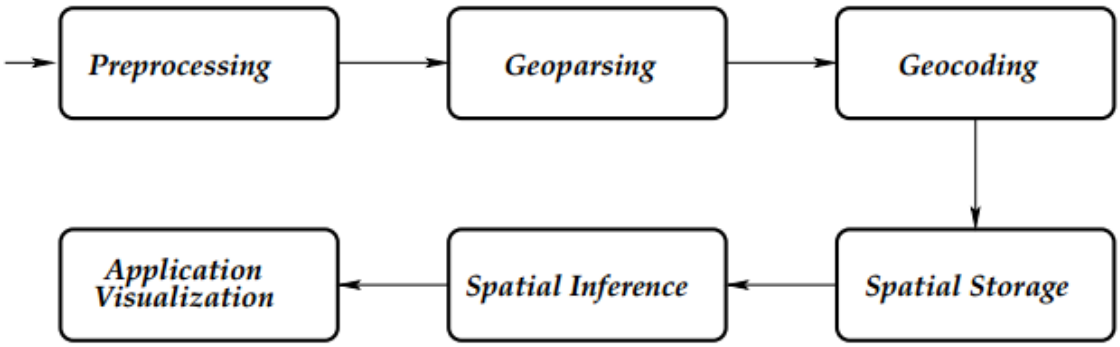
2.5.2 Paikanilmaisujen erottaminen

Geoparsing on jäsentämisprosessi, jossa tekstistä pyritään erottamaan paikkaan viittaavia sanoja (Location Indicative Words, LIW) tavallisista sanoista kuten ”tänään” ja ”ainakin”. Geoparsingilla pyritään nopeuttamaan geokoodausprosessia ja estämään tavallisten sanojen virheellinen paikantaminen (Han ym., 2014). Geoparsing tunnetaan myös nimillä geotagging, georecognition ja toponym recognition. Geoparsingin avulla erotetut paikannimet voidaan geokoodata, jotta niille saadaan koordinaatit. (Leidner & Lieberman, 2011) Geoparsing on osa laajempaa kokonaisuutta nimeltä Named Entity Regcognition (NER), jossa pyritään luokittelemaan tekstin osat esimerkiksi henkilöiden nimiin, paikkoihin, organisaatioihin, aikamääreisiin, suureisiin ja rahayksiköihin. (Freire ym., 2011) Named entity tarkoittaa nimettyä kohdetta, eli yhdestä tai useammasta sanasta koostuvaa ilmausta, joka viittaa tiettyyn kohteeseen (Tieteen termipankki, 2017). NER puolestaan on osa tiedon eristämistä, jossa pyritään automaattisella prosessilla jäsentelemään jäsentelemätöntä aineistoa, jotta sitä voidaan paremmin käsitellä laskentamenetelmillä. (Freire ym., 2011)

Geoparsingin, ja niin ollen myös geokoodauksen, isoin haaste on sanojen monimerkityksellisyys. Paikannimien tunnistamisen kohdalla ongelmat voidaan luokitella kahteen luokkaan, geo/non-geo ja geo/geo ongelmiin. Geo/non-geo ongelma ilmenee kun paikannimellä on toinen merkitys, joka ei viittaa paikkaan, esimerkiksi Georgialla voidaan viitata joko paikkaan tai henkilöön. Geo/geo ongelma puolestaan syntyy, kun useilla paikoilla on sama nimi. Haetulla geokoodauksen tasolla, eli käytännössä vertailuaineistolla, on merkittävä vaikutus näiden virheiden määrään. Jos halutaan löytää paikannimet maa-tasolla, on väärintulkintojen riski paljon pienempi, kuin jos mukaan otetaan suuri määrä kaupunkeja, jolloin monimerkityksellisiä sanoja esiintyy runsaasti. Toinen merkittävä haaste ovat, samoin kuin osoitteiden kohdalla, kirjoitusvirheet. (Leidner & Lieberman, 2011; Freire ym., 2011) Kun tekstiä verrataan myös maiden nimiin aiheuttavat kuitenkin kielet ja kansallisuudet sekaannuksia. Esimerkiksi tässä työssä käytetyllä pakkoruotsiaineistolla se olisi johtanut monien twiittien virheelliseen paikantamiseen Ruotsiin. Toisaalta voi olla myös toivottavaa että esim. tsekki (henkilö) pystytään yhdistämään Tsekkiin. Monimerkityksellisten sanojen osalta on mahdotonta tietää, kumpaa tekstissä tarkoitetaan pelkästään yhtä sanaa tarkastelemalla, vaan tällöin on katsottava koko lausetta jotta voidaan päätellä, onko kyseessä paikka vai ei. Kun ohjelma on onnistunut paikantamaan esimerkiksi merkkijonot ”soini” ja ”levi” on tarkistettava tekstin asianyhteydestä, onko kyseessä henkilö vai paikka.

Paikkoihin viittaavat ilmaisut voivat esiintyä teksteissä monissa eri muodoissa. Tyypillisimpiä lienee maan tai kaupungin nimet, kuten Suomi ja Espoo, mutta ne voivat olla myös katuja, katujen risteyksiä, rakennuksia ja kaupunginosia. Joissain tapauksissa myös puistot,

joet ja vuoret voivat olla kiinnostavia kohteita. Lisäksi voidaan olla kiinnostuneita tarkemmista määritelmistä kuten ”etelä-Ranska” tai ”30 km Espoon pohjoispuolella”. Geoparsing on yleensä osa isompaa kokonaisuutta, josta kuvassa 3 on esimerkki. Preprocessing, eli alkuvalmisteluvaihe saattaa tekstin laadusta riippuen sisältää vaihtelevan määrän vaiheita pelkästä tekstin osan valinnasta, layoutin analysoinnista kirjoitusvirheiden tunnistamiseen. (Leidner & Lieberman, 2011) Jäljempänä käydään läpi esimerkkinä mitä alkuvalmisteluita Zhang ja Gelernter (2013) käyttävät omassa prosessissaan.



Kuva 3. Geoparsing osana isompaa kokonaisuutta (Leidner & Lieberman, 2011)

Geokoodausosion jälkeen voidaan tieto tallentaa muotoon joka mahdollistaa nopean hakemisen sijainnin perusteella, esim. kaikki kohteet tietyn alueen sisällä, tai tehdä loogisia päättelyitä. Esimerkiksi jos A on B:n etelä puolella ja B on C:n eteläpuolella, niin silloin A on C:n eteläpuolella. Tyypillisesti paikkatietoanalyyseissä halutaan vielä lopuksi visualisoida tulos esim. pisteinä tai klustereina kartalla. (Leidner & Lieberman, 2011)

2.5.3 Geoparsing-menetelmät

Geoparsingiin käytettävät metodit voidaan jakaa kolmeen eri luokkaan. Gazetteer based lookup perustuu tekstin yhdistämiseen vertailuaineistoon, eli teksti käydään läpi sana sanalta ja verrataan vertailuaineistoon, joka sisältää paikannimiä. Tässä tapauksessa saatetaan tarvita erilliskäsittelyä useista sanoista koostuville nimille, kuten New York City, koska nimen pilkkominen osiin, johtaa epätoivottuun tulokseen. (Leidner & Lieberman, 2011) Toinen luokka on sääntöihin perustuva geoparsing, jossa käytetään tietyn kielistä sääntöjoukkoa, jonka perusteella pyritään ratkaisemaan, onko kyseessä paikannimi vai ei. Esimerkiksi ”city of” ilmaisun jälkeinen sana on paikka. Pääsääntöisesti tähän käytetään säännöllisiä lausekkeita (regular expressions), kuten FiniteState Automata (FSA) tai Context-Free Grammars (CFG). (Leidner & Lieberman, 2011) Ruotsinkieliselle tekstille säännöt voivat olla, että jokaisen paikkaan viittaavan preposition jälkeinen sana otetaan mukaan geokoodausvaiheeseen. Tällaisia ovat esimerkiksi i, till, mot ja från. Tällöin lauseesta ”Vi är påväg till Helsingfors” otetaan mukaan sana Helsingfors, joka voidaan onnistuneesti yhdistää vertailuaineistoon. Toisaalta lauseesta ”Jag har matteprov i skolan i morgon” otetaan mukaan ”skolan” ja ”morgon”, jotka eivät kumpikaan viittaa tunnistettavaan paikkaan. Geokoodattavien sanojen määrä kuitenkin vähenee huomattavasti verrattuna siihen, että kaikki sanat yritetään paikantaa. Samankaltainen logiikka pätee myös englanninkielisiin teksteihin, vaikkakin eri prepositioilla. Suomenkieli tuottaa kuitenkin sääntöihin perustuvalle geoparsingille ongelmia, sillä paikkaa ei ilmaista prepositioilla vaan päätteillä, esim. Tampereella, Kemissä, Turusta, Helsinkiin, Kaustiselle jne. Erilaisia taivutuksia on lukuisia ja vaikka pystytään tekemään säännöt sille mitkä kaikki kirjainyhdistelmät ovat paikkaan viittaavia päätteitä, niin

pelkkä niiden poistaminen ei kaikissa tapauksissa riitä, jotta paikannimi saataisiin perusmuotoon. Esimerkiksi merkkijonojen ”Kaustiselle” ja ”Kaustinen” välille jää eroa vielä kolmen kirjaimen verran, vaikka ”lle” pääte poistettaisiin. Tällöin on vielä löydettävä sopiva rajaarvo yhdistämisalgoritmille, jotta se osaa yhdistää tämänkaltaiset sanat. Jos sääntöihin perustuvaa geoparsingia käytetään aineistolle jossa esiintyy tekstejä monilla eri kielillä, on järjestelmässä oltava jokaiselle eri kielelle omat säännöt ja lisäksi tarvitaan tieto siitä millä kielellä mikäkin teksti on kirjoitettu tai vaihtoehtoisesti twiitit on ensin käännettävä samalle kielelle.

Kolmas geoparsing luokka on koneoppimiseen perustuvat menetelmät. Koneoppiminen on tekoälyn osa-alue, joka geokoodauksessa perustuu algoritmeihin, joille opetetaan sanojen ja paikkojen välinen yhteys käyttämällä opetusdataa. Tällöin on mahdollista yhdistää myös muita sanoja kuin paikannimiä paikkoihin. Koneoppimista käytetään useissa geokoodausta käsittelevissä tutkimuksissa (Zhang ja Gelernter, 2014; Sidkar ja Gambäck, 2016) ja sillä on aikaisemmissa tutkimuksissa (esim. Zhang ja Gelernter, 2014) saatu hyviä tuloksia. Järjestelmä opetetaan tekemään oikeita ratkaisuja eri tilanteissa käyttämällä opetusaineistoa, eli niin kutsuttua gold dataa. Opetusaineisto luodaan käymällä se manuaalisesti läpi ja merkitsemällä siinä esiintyvät paikannimet. Prosessi on työläs ja aikaa vievä, minkä vuoksi käytettävät opetusaineistot ovat yleensä aika pieniä verrattuna analysoitavan datan määrään. On olemassa joitain valmiita opetusaineistoja sanojen luokittelua varten, mutta aineistot jotka soveltuvat hyvin juuri geoparsingiin ovat harvassa. Työläyden lisäksi yksi syy tähän on, että aineistojen käyttö on usein rajoitettua, mikä estää niiden jakamisen tutkijoiden kesken. (Leidner & Lieberman, 2011) Twiittien geokoodaukseen käytettävän koneoppimismenetelmän on syytä olla helposti uudelleen opetettava johtuen twiittien alati muuttuvasta sisällöstä ja aihepiireistä. (Han ym., 2014)

Nykyään NER suoritetaan pääasiassa koneoppimismenetelmiä hyödyntämällä ja sen katsotaan olevan tyypillinen ongelma, johon koneoppimista voidaan hyödyntää. Sillä voidaankin päästä lähelle ihmisen tekemän työn tasoa englanninkielisillä virallisilla teksteillä. Yksi syy koneoppimisen yhä lisääntyvään käyttöön on sen muunneltavuus erilaisia tilanteita varten. Tästä huolimatta myös alkuperäisiä menetelmiä, eli sanojen vertailuun ja sääntöihin perustuvia menetelmiä käytetään edelleen. (Freire ym., 2011) Twiittien teksti ei noudata kielioppia samalla tavalla kuin esimerkiksi uutiset ja lisäksi tekstit ovat lyhyitä, mikä tekee paikannimien erittelemisestä haastavampaa. Kun teksti, josta paikannimiä ollaan erottelemassa, on useilla eri kielillä, vaikeutuu tehtävä entisestään. Useissa tutkimuksissa on yritetty kehittää menetelmiä paikannimien erottamiseen twiiteistä, mutta se on edelleen haastava tehtävä twiittien vaihtelevasta kirjoitusasusta johtuen (Sidkar ja Gambäck, 2016). Zhangin ja Gelernterin (2013) mukaan tavallisin tapa on ensin kääntää kaikki tekstit ensin samalle kielelle, tyypillisesti englanniksi, ja vasta sen jälkeen suorittaa geoparsing. Ainakin espanjasta, arabiasta ja swahilista englanniksi käännetyillä teksteillä, tämän on todettu toimivan.

Ennen geoparsingin varsinaista suorittamista tehdään pääsääntöisesti enemmän alkuvalmisteluja kuin tavallisessa geokoodattavan syötteen siivoamisessa. Zhangin ja Gelernterin (2013) mallissa teksti jaetaan sanoihin ja lausekkeisiin twiiteille kehitetyn tokenisointiprosessin avulla, jonka jälkeen teksti muutetaan pieniksi kirjaimiksi ja poistetaan ylimääräiset erikoismerkit. Useimmissa twitterissä paljon käytetyissä kielissä, esim. englannissa ja ranskassa, tokenisointi voidaan tehdä välilyöntien ja välimerkkien perusteella. Japanissa ja kiinassa näin ei kuitenkaan voida tehdä, vaan tokenisointiprosessi on paljon haastavampi. (Han

ym., 2014) Sanoihin jakamisen jälkeen tekstistä suodatetaan pois kaikki tavalliset sanat käyttämällä sanastoa, josta on poistettu kaikki paikkaan viittaavat sanat, jolloin saadaan pienennettyä vertailtavien sanojen määrää huomattavasti. Lisäksi tarkistetaan, mitkä sanat voidaan suoraan yhdistää vertailuaineistoon ja lopuille tehdään oikeinkirjoituksen tarkistus. Oikeinkirjoituksen tarkistuksessa sanat pyritään yhdistämään vertailuaineistoon epätarkalla vertailulla. Zhang ja Gelernter (2013) olettavat lyhyiden (alle neljän merkin) sanojen twiitissä todennäköisesti olevan lyhenteitä ja siksi niitä ei tarkisteta. Tämä saattaa kuitenkin johtaa siihen, että joitain paikkoja, joilla on lyhyt nimi jää huomioimatta. Yhdistämisalgoritmina Zhang ja Gelernter (2013) käyttävät sekä n-grammia että Levenshteinin etäisyyttä ja lisäksi he tarkastelevat kuinka usein paikannimi esiintyy vertailuaineistosta saatujen ehdokkaiden joukossa, olettaen että mitä useammin se esiintyy, sitä todennäköisemmin se on oikein. Kun mahdolliset kirjoitusvirheet on korjattu, suoritetaan perusmuotoistaminen (lemmatization), jossa poistetaan sanojen taivutukset ja muutetaan sanat niiden perusmuotoon, esimerkiksi ”puhuvat” muutetaan muotoon ”puhua”. Tämän lisäksi tunnistetaan vielä lausejäsenet (part-of-speech, POS), eli lisätään sanoille tieto siitä, onko kyseessä substantiivi vai verbi jne. Varsinaiseen jäsentely vaiheeseen Zhang ja Gelernter (2013) käyttävät kolmea eri tyyppistä geoparseria. Named Location Parser pyrkii yhdistämään nimetyt kohteet vertailuaineistoon joko tarkalla tai epätarkalla vertailulla. Suodatusprosessi estää useat virheelliset yhdistämiset, mutta tilanteet, joissa vain osa nimestä on löydetty tai nimi sisältää useita tavallisia sanoja jotka suodatetaan pois, kuten ”Blue Farm”, aiheuttavat ongelmia. Zhangin ja Gelernterin Named Entity Recognition perustuu koneopetukseen, jota varten he käyttivät manuaalisesti paikannettua yli 3000 twiitin opetusaineistoa. Katu ja rakennus parserit perustuvat sääntöihin ja sanaluokkiin, esim. etsimällä tiehen viittaavia sanoja joita seuraa substantiivi.

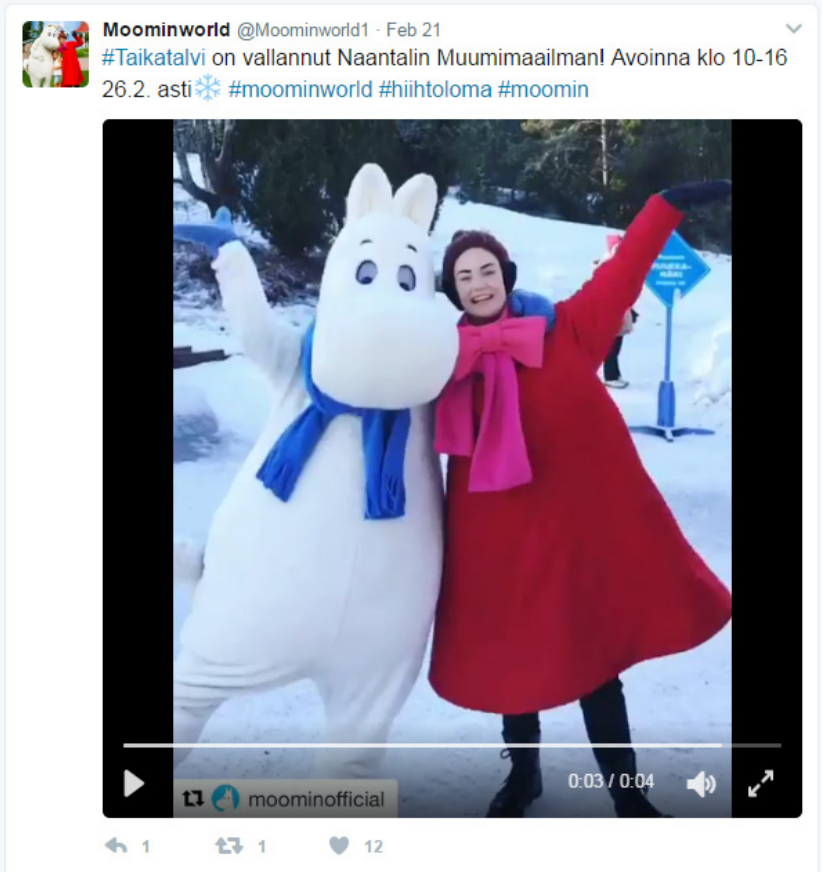
Sidkar ja Gambäck (2016) käyttävät työssään useita eri ominaisuuksia tekstin luokitteluun. Yksi niistä on hukkasanalistaa (stop-words). Hukkasanat ovat kielessä yleisiä sanoja, jotka esiintyvät melkein kaikissa dokumenteissa ja prosessin nopeuttamiseksi tai tilan säästämiseksi ne usein ohitetaan tekstinkäsittelyssä (Tieteen termipankki, 2017). Toinen on sanojen esiintymistiheys. Sanat jotka esiintyvät teksteissä harvoin ovat heidän mukaansa usein nimettyjä kohteita ja siksi jokaiselle sanalle merkitään tieto siitä ylittääkö sanan esiintymistiheys tietyn arvon. Muita tarkastettuja ominaisuuksia ovat mm. alkaako sana isolla kirjaimella ja sisältääkö se numeroita.

On olemassa useita patentoituja ja avoimen lähdekoodin geoparsereita. Tällaisia ovat esimerkiksi Thomson Reutersin OpenCalais, joka tosin ei toimi suomenkielisillä teksteillä, Apache:n OpenNLP, joka sisältää mm. moduulin nimettyjen kohteiden tunnistamista varten ja jota on mahdollista opettaa omalla opetusaineistolla, sekä Clavin ja LingPipe.

3 Twiittien geokoodaus käytännössä

3.1 Aineisto

Tutkimuksessa on käytetty kahta eri aineistoa. Ensimmäisen aineiston twiitit on kerätty ajanjaksolla 20.2-8.3.2017 ja hakukriteerinä on käytetty sanaa ”hiihtoloma”. Hiihtoloma valikoitui aiheeksi koska monet ihmiset matkustelevat hiihtolomalla ja sen myötä lisäävät twiitteihinsä mainintoja paikoista normaalia enemmän. Testiaineistossa on hyvä olla paikannimiä mainittuina, jotta saadaan selville, pystyykö kehitetty geokoodausjärjestelmä löytämään ne. Kaikki uudelleen twiittauset, eli retwiitit, poistettiin aineistosta, sillä ne kuvaavat oletettavasti alkuperäisen twiittaajan sijaintia ja tekemisiä eikä uudelleentwiittaajan. Twiittejä jäi karsinnan jälkeen jäljelle 911 kpl. Oletus paikkatietojen poikkeuksellisen suuresta määrästä osoittautui oikeaksi. Peräti 24 twiittiä, eli 2,6 % kaikista, sisälsi käyttäjän mobiililaitteesta saadut koordinaatit, mikä on huomattavasti aikaisempien tutkimusten ilmoittamaa keskimäärää enemmän. Lisäksi 110 kpl, eli 12 %, sisälsi käyttäjän twiitille ilmoittaman paikan tekstimuodossa. Jotta geokoodauksen tulosten laatua voitaisiin arvioida, käytiin kaikki twiitit yksitellen läpi ja lisättiin uuteen sarakkeeseen tunnistetut paikannimet. Paikannimiin ei laskettu maita tai laajoja viittauksia, kuten Etelä-Suomi, sillä niitä ei käytetystä vertailuaineistosta löydy. Kaikki tiettyyn paikkakuntaan viittaavat nimet otettiin kuitenkin mukaan, riippumatta siitä löytyykö paikka vertailuaineistosta. Tällä tavalla vertailuaineistosta puuttuvat paikat eivät heikennä tuloksia liikaa, mutta toisaalta voidaan huomata, minkälaisia puutteita siinä esiintyy. Yhteensä 271 twiitin tekstistä löytyi paikannimi jossain muodossa.



Kuva 4. Esimerkki twiitistä, joka onnistuttiin paikantamaan.

Kuvassa 4 nähdään esimerkki twiitistä, jonka onnistuttiin paikantamaan tässä kehitetyillä työkaluilla. Twiitin tekstissä esiintyy sana Naantalin, joka on niin lähellä perusmuotoa Naantali, että vertailualgoritmit osasivat yhdistää ne.

Toinen tutkimuksessa käytettävä aineisto on kerätty ajanjaksolla 25.12.2016 - 3.1.2017 ja siinä kaikissa twiiteissä esiintyy sana pakkoruotsi. Tämän testiaineiston on tarkoitus kuvastaa tapausta, jossa twiiteissä ei esiinny poikkeuksellisen paljon paikannimiä, vaan ne käsittelevät pääasiassa muita asioita. Tästäkin aineistosta poistettiin kaikki uudelleentwiittaukset. Lisäksi aineistossa esiintyi lukuisia twiittejä, joissa oli identtinen teksti ja sama lähettäjä, vaikka niitä ei oltukaan merkattu uudelleentwiitatuiksi. Myös näistä poistettiin kaksoiskappaleet, mutta ensimmäiseksi lähetetty twiitti säilytettiin. Jäljelle jäi yhteensä 1051 twiittiä. Yksikään tämän poiminnan twiitti ei sisältänyt koordinaatteja, mikä hyvin kuvastaa tarvetta käyttää muita tapoja twiittien sijainnin määrittämiseen. Kun koordinaatteja ei ole, on ensisijainen tapa paikallistaa twiitti place-kentän arvojen perusteella. Koeaineistosta löytyy 29 twiittiä, jolle on ilmoitettu sijainti käyttämällä paikkakunnan nimeä, mikä on vain 2,8 % prosenttia kaikista twiiteistä. Pakkoruotsiaineistosta löytyi ainoastaan kymmenen twiittiä, jonka tekstissä mainitaan jokin paikkakunta, joten tämän aineiston kohdalla pääasialliseksi paikannuskeinoksi jäi twiittajan kotipaikan paikantaminen. Molempien aineistojen sisältö on esitetty taulukossa 3. Tilastollisen tarkastelun kannalta olisi parempi, jos twiittien määrä olisi huomattavasti suurempi, esim. 10 000 kpl. Tässä on kuitenkin jäljitelty todellista hakutilannetta, jolloin tutkittua aihetta koskevien twiittien määrä on rajallinen.

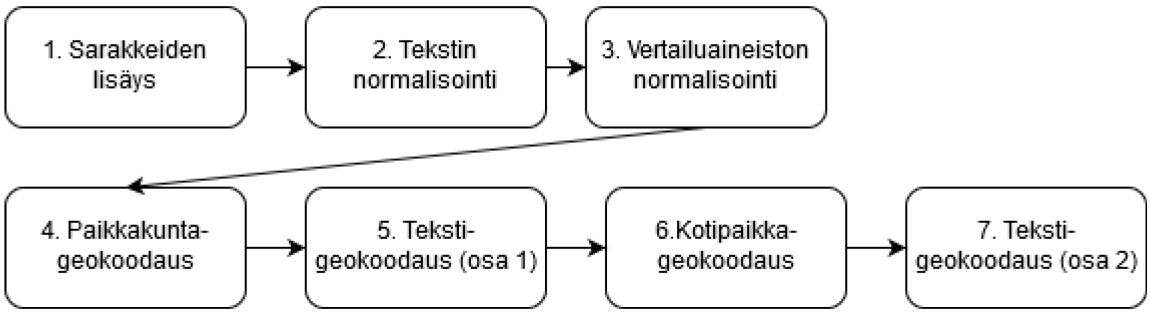
Taulukko 3. Testiaineistojen tunnusluvut

	Hiihtoloma		Pakkoruotsi	
	kpl	%	kpl	%
Twiitit	911		1051	
Koordinaatit	24	2,6	0	0
Paikka	110	12,1	29	2,8
Kotipaikka	676	74,2	707	67,3
Paikannimi mainittu tekstissä	271	29,7	10	1,0

Vertailuaineistona tässä työssä on käytetty GeoNamesista (www.geonames.org) ladattua aineistoa, joka sisältää noin 150 000 yli 1000 asukkaan kaupunkia ympäri maailman. Geonames on yleisesti käytetty vertailuaineisto paikannimien tunnistamista ja geokoodausta varten, johtuen juuri sen maailmanlaajuisesta kattavuudesta. Geonamesia ovat tutkimuksissaan käyttäneet mm. Freire ym. (2011), Zhang ja Gelernter (2013), Han ym. (2014) sekä Alex ym. (2016). Aineisto sisältää kaupunkien nimien lisäksi niiden koordinaatit sekä valtaosalle kaupungeista myös vaihtoehtoisia nimiä useilla eri kielillä ja vaihtelevilla kirjoitusasuilla. Tämän lisäksi on olemassa väkimäärä ja kaupungin tyyppi, sekä muuta tietoa, jota tässä tutkimuksessa ei käytetä. Koordinaatit ovat WGS84 koordinaattijärjestelmässä ja koska ylimääräisiä muunnoksia ei tehdä on geokoodauksen tulos samassa koordinaattijärjestelmässä. Kaupunkiaineistoon on yhdistetty Geonamesin CountryInfo aineisto, jotta jokaiselle kaupungille saadaan maan nimi yhteen sarakkeeseen. Maille on tällä hetkellä Geonamesissa vain yksi nimi, eikä ollenkaan vaihtoehtoja eri kielillä. Tästä syystä tätä työtä varten vertailuaineistoon on manuaalisesti lisätty muutamia syötteessä esiintyviä vaihtoehtoisia maiden nimiä.

3.2 Twittergeokoodauksen toteutus

Geokoodaus suoritettiin käyttämällä Esrin ArcMap 10.3 ohjelmaa ja siihen tätä tutkimusta varten kehitettyjä työkaluja. Työkalujen kehittämisessä käytettiin ArcMapin omien työkalujen lisäksi Python ohjelmointikielellä itse tehtyjä skriptejä, jotka on mahdollista sisällyttää osaksi ArcMapin työkaluja. Kehitetty työkalupaketti sisältää seitsemän eri työkalua, jotka on esitetty kuvassa 5. Kolmea ensimmäistä käytetään syötteen sekä vertailuaineiston valmisteluun ja normalisointiin ennen varsinaista geokoodausta. Työkalut 4-7 ovat varsinaista geokoodausta varten ja osalle niistä on olemassa useita vaihtoehtoja, jotka käyttävät eri yhdistämisalgoritmeja.



Kuva 5. Työkalupaketin sisältö ja oletettu suorittamisjärjestys

Vertailuaineiston tietotyyppinä käytetään Pythonin sanakirjaa (dictionary), joka koostuu avain-arvo pareista. Avaimena toimii monikko (tuple) joka sisältää kaupungin nimen ja id-numeron. Id numeron käyttö on välttämätöntä johtuen aineistossa esiintyvistä samannimisistä kaupungeista, sillä kaikkien sanakirjan avainten on oltava erilaisia. Avainta vastaavana arvona on toinen monikko, joka sisältää kaikki muiden kenttien arvot, kuten vaihtoehtoiset nimet, koordinaatit jne. erotettuna toisistaan pilkulla. Sanakirja on kätevä, sillä kun saadaan yhdistettyä merkkijono ja kaupunki, niin avaimen avulla voidaan helposti hakea loput tiedot kirjoitettavaksi twiitille.

3.2.1 Twiittien ja vertailuaineiston normalisointi ja geoparsing

Ensimmäisellä työkalulla syötteen, eli twiitit, sisältävään taulukkoon lisätään tarvittavia sarakkeita koordinaateille, geokoodauksen tulokselle sekä vaihtoehtoisille sijainneille. Lisäksi twiitin teksti ja käyttäjän ilmoittama kotipaikka kopioidaan uuteen kenttään, jotta ne säilyvät tallella, vaikka niitä seuraavaksi karsitaan. Normalisointi toteutetaan korvaamismenetelmällä sen helppokäyttöisyydestä johtuen. Sekä syöte että vertailuaineisto muutetaan sisältämään ainoastaan pieniä kirjaimia, jotta niistä saadaan vertailukelpoisia riippumatta siitä, onko paikannimet twiiteissä kirjoitettu pienellä vai isolla alkukirjaimella. Twiittien tekstikentästä ja käyttäjätilin kotipaikkaa kuvaavasta kentästä poistetaan lisäksi erikoismerkit jotka saattavat haitata vertailua, kuten # ? ! @, sekä kaikki numerot.

Twiittien normalisointityökalulla suoritetaan lisäksi tekstin geoparsing, eli erotellaan kaikki paikkaan viittaavat ilmaisut. Koska sekä koneoppiminen että sääntöihin perustuva geoparsing, joita käsiteltiin kappaleeessa 2.5.3, todettiin liian haastaviksi toteuttaa tämän työn puitteissa päädyttiin käyttämään sanastoa suodatuksena, jotta saadaan tavalliset sanat pois vertailtavasta tekstistä. Tämä tuottaa haasteita etenkin paljon taivutuksia sisältävälle kielelle, kuten suomelle. Koska sanasto sisältää vain sanojen perusmuodot on teksti ensin pystyttävä perusmuotoistamaan. Perusmuotoistamiseen on olemassa valmiiksi kehitettyjä sovelluksia, mutta niiden käytössä esteeksi muodostuu niiden soveltuvuus pääasiassa englanninkieliselle

tekstille. Zhangin ja Gelernterin (2013) mukaan on yleistä käyttää ensin kääntäjää, joka kääntää kaikki twiitit englanniksi. Käännetylle tekstille voidaan suorittaa perusmuotoistaminen ja sen jälkeen suodattaa sanastosta löytyvät sanat pois. Tällöin jäljelle jäävät nimet, sekä kaikki sanat joita kääntäjä ei onnistunut kääntämään, kuten puhekieli ja murresanat.

Yhtenä ratkaisuvaihtoehtona kokeiltiin oman rajauslistan käyttöä. Rajauslistaan lisättäisiin tällöin vain ne sanat, jotka on todettu paikannettavan väärin, eikä siitä tulisi ihan yhtä massiivinen kuin kokonainen sanasto. Tällainen rajauslista on kuitenkin ongelmallinen käyttää. Rajauslistaa ei voida tehdä osaksi ohjelmaa, sillä käyttäjän on pystyttävä määrittämään mitä sanoja listaan halutaan laittaa. Sanamäärä nousee nopeasti niin suureksi, että on vaikea pysyä selvillä mitkä sanat ovat mukana ja mitkä eivät. Tämän seurauksena listalle saattaa päätyä sanoja, joita jossain toisessa tapauksessa haluttaisiinkin käyttää paikantamiseen. Lisäksi rajattavat sanat vaihtelevat tekstin kielen mukaan ja niitä sanoja jotka rajataan suomenkielisestä tekstistä pois ei todennäköisesti haluta rajata englanninkielisestä tekstistä.

Väärin paikannettujen sanojen vähentämiseksi päädyttiin lopulta käyttämään hukkasanalistaa sen mukaan, millä kielellä twiitti on kirjoitettu. Twitter pyrkii tunnistamaan jokaisen twiitin kielen ja ilmoittaa sen osana metatietoja. Tätä tietoa käytetään tässä työssä määrittämään minkä kielistä hukkasanalistaa kulloinkin tulisi käyttää. Twitterin kielentunnistus ei ole sataprosenttinen, sillä sitä vaikeuttaa mm. se, että monet käyttäjät sekoittavat useita kieliä samaan twiittiin. Lisäksi vain parin sanan twiitit ja murteiden käyttö hankaloittavat kielen tunnistamista. Mikäli kieltä ei ole pystytty tunnistamaan tai sitä ei löydy ohjelman käytössä olevista hukkasanalistoista käytetään oletuskielenä suomea. Suomea siksi koska tutkimuksessa käytettävä aineisto on pääasiassa suomenkielistä. Jos aineiston twiitit jakautuisivat laajemmalle alueelle ympäri maailman olisi luonnollista käyttää oletuskielenä englantia. Hukkasanalistat on otettu sivustolta www.ranks.nl, joka tarjoaa hukkasanalistoja useilla eri kielillä. Twiittien teksteistä poistetaan jo normalisointivaiheessa kaikki hukkasanalistassa esiintyvät sanat, jotta varsinaisessa tekstin geokoodausvaiheessa olisi vähemmän prosessoitavaa.

Valtaosa virheellisesti paikannetuista tavallisista sanoista ovat lyhyitä sanoja, kuten ”lue” tai lyhenteitä kuten ”rkp”. Ongelmaa saataisiin vähennettyä huomattavasti poistamalla kaikki korkeintaan neljän merkin sanat, kuten Zhang ja Gelernter (2013) tekivät omassa työssään. Haittapuolena tässä on kuitenkin se, että samalla poistuu mahdollisuus paikantaa korkeintaan neljän merkin pituiset paikannimet. Tällaisia paikannimiä esiintyy 150 000 kaupungin vertailuaineistossa 5300 kpl jo pelkästään virallisten nimien joukossa ottamatta huomioon vaihtoehtoisia nimiä. Kahden merkin pituisia nimiä (esim. Ii) esiintyy 55 kpl. Koska ne ovat kuitenkin kohtalaisen harvassa, päädyttiin tässä työssä vaihtoehtoon, jossa korkeintaan kahden merkin pituiset merkkijonot poistetaan jo normalisointivaiheessa.

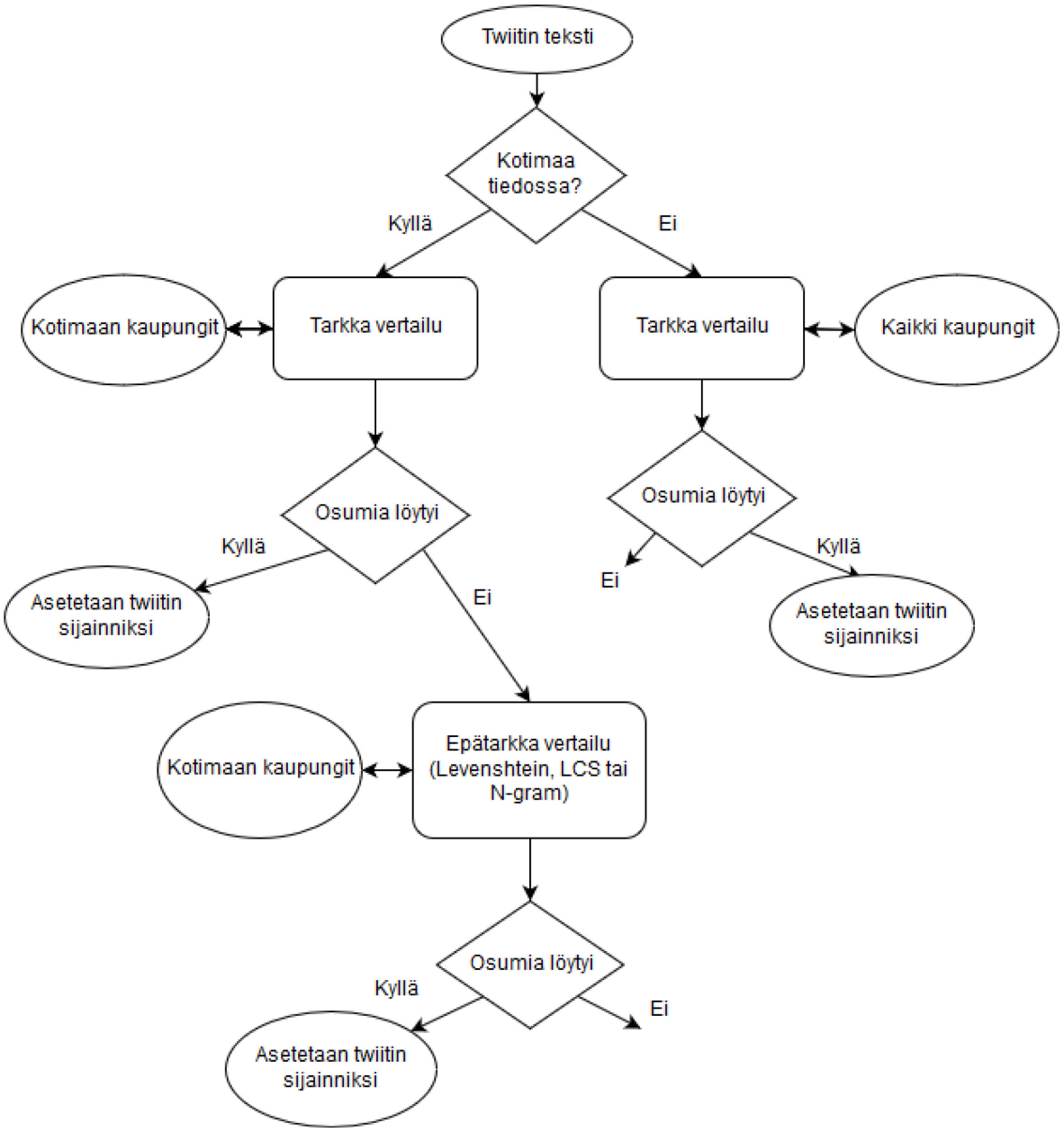
3.2.2 Geokoodaustyökalut

Työkalut 4-7 ovat varsinaista geokoodausta varten. Niistä työkaluille 5 ja 7 on olemassa kolme vaihtoehtoista työkalua, jotka käyttävät eri menetelmiä epätarkkaan vertailuun, mutta ovat muilta toiminnoiltaan samanlaiset. Twiitit jotka sisältävät käyttäjän laitteen antamat koordinaatit eivät vaadi geokoodausta, vaan ne voidaan sijoittaa suoraan kartalle. Lopuista twiiteistä käydään läpi ne, joille käyttäjä on antanut sijainniksi paikkakunnan. Tämä tieto on aineistossa automaattisesti muotoa ”paikkakunta, maa”, mikä tekee siitä suhteellisen helpon geokoodattavan. Haastetta aiheuttavat kuitenkin maat joissa on useita samannimisiä kaupunkeja. Geonamesista löytyy esimerkiksi 13 kpl Cambridge nimistä kaupunkia, joista 8 sijaitsee Yhdysvalloissa. Näistä kaupungeista ensimmäiseksi löydetty lisätään twiitin sijainniksi

ja loput seitsemän listataan vaihtoehtoisten sijaintien kenttään. Tällaisessa tilanteessa olisi jatkossa mahdollista kehittää järjestelmää niin, että sijainniksi laitettaisiin se paikka, jossa on Geonamesin mukaan eniten asukkaita, ja joka näin ollen on todennäköisimmin mainittu. Twiitin paikan kaupunkia verrataan vain ilmoitetun maan kaupunkeihin. Mikäli maata ei löytyy vertailuaineistosta, esimerkiksi jos maan nimi on ilmoitettu kirjainlyhenteenä, verrataan kaupunkia koko vertailuaineistoon. Tätä samaa periaatetta käytetään myös muiden kenttien perusteella tehtävissä vertailuissa. Paikkaa verrataan vertailuaineistoon tarkalla vertailulla, eli determininstisellä menetelmällä, jolloin osuma joko löytyy tai ei löydy.

Twiitit joita ei pystytä paikan perusteella paikantamaan pyritään paikantamaan varsinaisen tekstin avulla. Jotta virheellisesti paikannimeksi tulkittujen sanojen määrää saataisiin alas, päädyttiin ensin tarkastelemaan twittertilin omistajan ilmoittamaa kotipaikkaa. Jos tilin kotipaikaksi on merkitty Suomi tai jokin kaupunki Suomessa oletetaan, että twiitissä mahdollisesti esiintyvät paikannimet sijaitsevat Suomessa. Tämä oletus ei tietenkään pidä läheskään aina paikkansa, sillä twiitti voi olla lähetetty ulkomaanmatkan aikana tai twiittaaja voi olla kotimaassaan, mutta kirjoittaa Pariisissa sattuneesta terrori-iskusta. Herää kysymys, halutaanko saada selville twiittaajan senhetkinen sijainti, paikka jota twiitti käsittelee vai twiittajan kotipaikka? Tässä on päädytty ratkaisuun, jossa ensisijaisesti pyritään selvittämään twiittauspaikka. Jollei sitä löydy niin tekstissä mainittu paikka ja viimeisenä vaihtoehtona twiittaajan kotipaikka, eli käytännössä kaikki yllä mainitut tietyssä järjestyksessä. Koska twiitin eri kenttien perusteella tehtävät geokoodaukset suoritetaan erillisillä työkaluilla, voidaan kuitenkin haluttaessa esimerkiksi suorittaa geokoodaus pelkästään kotipaikan perusteella. Vaikka twiittien rajaaminen kotimaahan aiheuttaa jonkin verran puutteita tuloksiin se koettiin kuitenkin paremmaksi vaihtoehdoksi kuin, että saadaan valtava määrä virheellisesti paikannettuja sanoja. Jotta kotimaata ei tarvitsisi etsiä moneen kertaan ja koska sitä saatetaan haluta käyttää tulosten analysoinnissa, tallennetaan löydetty kotimaa erilliseen sarakkeeseen. Kotimaan sijaan rajaavaksi tekijäksi harkittiin myös twiitin kielen käyttöä. Kielet ovat kuitenkin hankalia koska useissa maissa puhutaan monia kieliä ja englanninkielisiä twiittejä olisi käytännössä verrattava melkein koko maailman kattavaan vertailuaineistoon. Näin ollen tästä ajatuksesta luovuttiin ja twiitin kieltä käytetään ainoastaan edellä mainitun hukkasanalistan valintaan.

Twiitin sanat käydään läpi yksitellen ja jokaisen sanan kohdalla käydään läpi kuvan 6 esittämä prosessi. Ensin sanalle tehdään tarkka vertailu samalla tavalla kuten paikka-kentälle. Jos kotimaa on tiedossa, verrataan sanaa vain sen maan kaupunkeihin, muussa tapauksessa koko vertailuaineistoon. Jos tarkkaa vastinetta ei löydy suoritetaan todennäköisyyspohjainen, eli epätarkka vertailu, käytössä olevalla yhdistämisalgoritmilla. Yhdistämisalgoritmille määritellään tietty raja-arvo, joka määrittää kuinka hyvin sanan on täsmättävä vertailuaineiston paikkaan, jotta se voidaan määrittää osumaksi. Samalle sanalle voi löytyä useita osumia ja algoritmi palauttaa ne kaikki, sekä vertailuparin samanlaisuusarvon. Epätarkka vertailu päädyttiin jakamaan kahteen osaan, sillä ilmeni, että epätarkka vertailu joka suoritetaan koko vertailuaineistoon, kun tiedossa ei ole twittertilin kotimaata, tuottaa enemmän virheellisisä kuin oikeita vastauksia ja vie lisäksi paljon aikaa suhteessa työkalun suorittamiin muihin vaiheisiin. Näin ollen ensimmäinen tekstingeokoodaustyökalu suorittaa epätarkan vertailun vain niiden twiittien sanoille, jotka pystytään rajaamaan tiettyyn maahan. Tämän jälkeen geokoodauksen suorittaja pystyy itse valitsemaan, haluaako hän, että epätarkka vertailu suoritetaan lopuille twiiteille koko vertailuaineistoa käyttäen vai ei.



Kuva 6. Tekstigeokoodauksen ensimmäisen osan prosessi

Kaikissa vertailuissa sanaa verrataan ensin vertailuaineistossa ilmoitettuun ”viralliseen” kaupungin nimeen ja sen jälkeen kaikkiin vaihtoehtoisiin nimiin yksitellen. Prosessin nopeuttamiseksi oletetaan, että sana jonka pituus eroaa huomattavasti kaupunginnimen pituudesta ei voi tuottaa tarpeeksi hyvää vertailutulosta, joten niiden samanlaisuutta ei edes kannata laskea, raja-arvona käytetään 3-4. Toinen oletus joka koskee epätarkkaa vertailua on, että sanan ensimmäinen kirjain on oikein, eli käyttäjä ei ole kirjoittanut ”elsinki” tarkoittaessaan Helsinkiä. Tämäkään oletus ei tietenkään aina pidä paikkaansa, mutta se nopeuttaa ohjelman suoritusta pienehköllä aineistolla jopa kolmannekseen alkuperäisestä, joten se päätettiin kuitenkin tehdä. Epätarkka vertailualgoritmi palauttaa jokaiselle vertailuparille arvon, joka kuvaa sanojen samanlaisuutta. Kaikki kaupungit joiden kohdalla vertailuarvo ylittää tietyn rajan lisätään vaihtoehtoisiksi sijainneiksi kyseiselle twiitille. Kun twiitin kaikki sanat on käyty läpi, valitaan kaikista vaihtoehdoista se joka parhaiten vastaa jotain twiitin sanaa ja sijoitetaan twiitti sinne. Loput sanat ilmoitetaan vaihtoehtoina, jotta käyttäjä voi halutessaan

vaihtaa sijainnin johonkin niistä. Koska sanoille joille löytyi tarkka vastine ei suoriteta epätarkkaa vertailua, syntyy mahdollisuus virheisiin, mutta samalla se nopeuttaa suoritusta verrattuna siihen, että kaikille sanoille tehtäisiin myös epätarkka vertailu. Virhe on mahdollinen, jos paikannimi on kirjoitettu väärin ja se sen vuoksi täsmää tarkasti johonkin toiseen paikkaan, kuin mihin sen kuuluisi ja oikea paikka ei tule edes vaihtoehtoihin, koska epätarkkaa vertailua ei kyseiselle sanalle suoriteta.

Epätarkkaa vertailua varten käytetään kolmea eri menetelmää, jotka ovat luvussa 2.2 esitellyt Levenshtein distance, LCS ja n-grammi. Eri menetelmiä käytetään, jotta niiden tuloksia voidaan verrata ja sen myötä mahdollistaa parhaiten soveltuvan menetelmän valinta. Longest common subsequence valittiin käytettäväksi Longest common substring sijaan, koska jälkimmäinen ei sovellu yhtä hyvin kirjoitusvirheitä sisältävien sanojen yhdistelyyn. N-grammin ja s-grammin osalta valinta osui n-grammiin, koska siihen löytyi valmis Pythonilla toteutettu algoritmi ja sen toiminnaltaan yksinkertaisempi. Myös muita yhdistämisalgoritmeja löytyy internetistä valmiiksi Pythonkielellä toteutettuna. Valmiita algoritmeja on tässä työssä käytetty, jotta niitä ei tarvitsisi kehittää uudelleen. Tästä johtuen ei ole kuitenkaan varmaa onko kyseiset algoritmit varmasti toteutettu optimaalisella tavalla tätä tarkoitusta varten. Koska ne ovat yleisesti käytettyjä oletetaan kuitenkin niiden olevan tarpeeksi hyviä. Yhdistämisalgoritmit palauttavat arvon, joka kuvaa verrattavien sanojen samanlaisuutta. Levenshteinin menetelmä palauttaa 0 jos sanat ovat identtiset ja 1 jos ne ovat täysin erilaiset, eli käytännössä arvo kuvaa pikemminkin sanojen erilaisuutta kuin samanlaisuutta. Jotta eri menetelmiä olisi helpompi verrata toisiinsa on LCS ja n-grammi normalisoitu palauttamaan arvoja samalla välillä. LCS-algoritmi palauttaa sanojen pisimmän yhteisen merkkijonon pituuden. Vertailukelpoisen arvon saamiseksi pituus jaetaan sanojen keskipituudella. Jakajaksi kokeiltiin myös vertailuaineiston kaupunginnimen pituutta, koska tällä tavalla päätteet paikannimen lopussa eivät saisi vertailussa painoarvoa, esimerkiksi jos verrataan merkkijonoja ”tampereella” ja ”tampere” saadaan tuloksena, että merkkijonot ovat samat. Tämä johtaa kuitenkin virheisiin, kun verrattavassa merkkijonossa esiintyy osana jokin paikannimi, esimerkiksi tapauksessa jossa verrataan merkkijonoja ”lillkyro” ja ”kyro”, saavat nämä yhtä hyvän samanlaisuusarvon kuin ”lillkyro” ja ”lillkyro”. Myös Ranzijn (2013) päätyi omassa työssään siihen, että keskiarvolla jakaminen on paras vaihtoehto, sillä se ei suosi liikaa kumpaakaan merkkijonoa. N-grammi algoritmi laskee merkkijonojen samanlaisuuden jakamalla yhteisten n-grammien määrän kaikkien erilaisten n-grammien lukumäärällä. Vaikka kaikki yhdistämisalgoritmit palauttava arvoja välillä 0-1, eivät ne ole suoraan vertailukelpoisia, ja siksi jokaiselle algoritmille selvitettiin paras raja-arvo kokeilemalla eri vaihtoehtoja.

Twiitit joita ei ole pystytty paikantamaan tekstin perusteella pyritään paikantamaan twiittaajan ilmoittaman kotipaikan perusteella. Kotipaikan paikannuksessa oletetaan käyttäjän ilmoittaneen kotipaikakseen kaupungin, maan tai molemmat. Tässä haussa hyödynnetään tekstinpaikannuksen yhteydessä aikaisemmin tallennettua kotimaata ja siksi tekstin paikannus on syytä suorittaa ensin. Kotipaikan paikannuksessa käytetään vain tarkkaa vertailua, sillä havaintojen mukaan useimmat käyttäjät jotka ovat kirjoittaneet paikannimen, ovat kirjoittaneet sen oikein. Tämä saattaa johtua siitä, että twittertilin luomisvaiheessa käyttäjä todennäköisemmin oikolukee tiedot kuin yksittäistä twiittiä lähettäessä. Epätarkkaa vertailua ei näin ollen tässä kohtaa tehdä, sillä se lisäisi todennäköisesti oikeiden tulosten määrää huomattavan vähän suhteessa sen vaatimaan aikaan. Kuten mainittua käyttäjä voi kirjoittaa kotipaikkakenttään mitä tahansa tai olla kokonaan kirjoittamatta, joten läheskään kaikille twii-

teille ei tälläkään menetelmällä saada minkäänlaista sijaintia. Halutessaan käyttäjä voi suorittaa toisen osan tekstigeokoodauksesta vasta viimeisenä vaihtoehtona, eli kotipaikkageokoodauksen jälkeen.

Tulosten analysointia ja visualisointia varten jokaiselle geokoodatulle twiitille annetaan tasoarvo, joka kertoo, minkä kentän perusteella se on paikannettu, sekä onko sille löytynyt muita vaihtoehtoisia sijainteja. Tasot ja niiden selitykset näkyvät taulukossa 4. Mainintа useita osumia tarkoittaa, että kyseisen twiitin tekstikentälle saatiin enemmän kuin yksi tulos. Sama sana on siis voitu paikantaa useaan paikkaan tai kyseisestä kentästä on voitu paikantaa useita sanoja.

Taulukko 4. Tasot ja niiden merkitykset

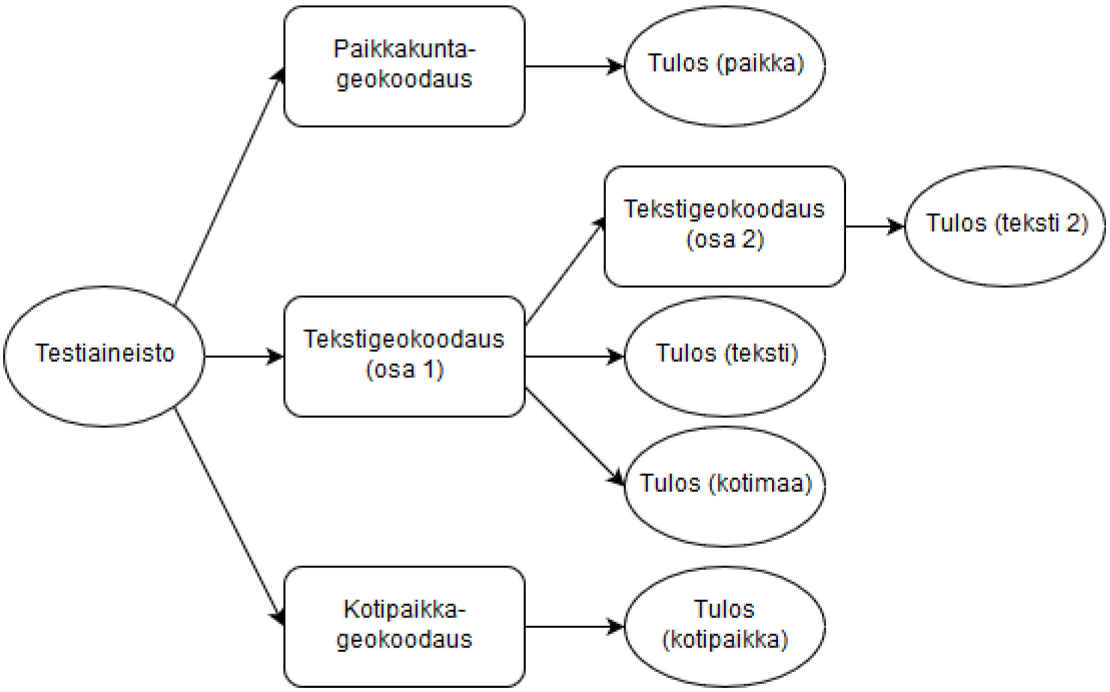
Taso	**Geookoodattava kenttä**	**Lisätieto**
1	Coordinates	Koordinaatit annettu
2	Place	Yksi osuma
25	Place	Useita osumia
3	Text	Tarkka vertailu. Yksi osuma
35	Text	Tarkka vertailu. Useita osumia
4	Text	Epätarkka vertailu. Kotimaa tiedossa. Yksi osuma
45	Text	Epätarkka vertailu. Kotimaa tiedossa. Useita osumia
5	AccountLocation	Yksi osuma
55	AccountLocation	Useita osumia
6	Text	Epätarkka vertailu. Kotimaa ei tiedossa. Yksi osuma
65	Text	Epätarkka vertailu. Kotimaa ei tiedossa. Useita osumia

Tason lisäksi paikannetuille twiiteille annetaan koordinaatit, jotka saadaan osuman tuottaneelta kaupungilta, sekä sana jonka perusteella twiitti paikannettiin. Jos osumia on useita, listataan loput osumat vaihtoehtokenttään. Mikäli osumat ovat löytyneet epätarkalla vertailulla ilmoitetaan samanlaisuusarvo ja mihin kaupunginnimeen osuman tuottanut vertailu tehtiin. Jos saadaan useita vaihtoehtoisia epätarkkoja osumia, valitaan paras vastaavuus twiitin sijainniksi. Näin ollen twiitti on voitu paikantaa virheellisesti tavallisen sanan perusteella, mutta vaihtoehdoista saattaa löytyä paikannimi ja sen koordinaatit.

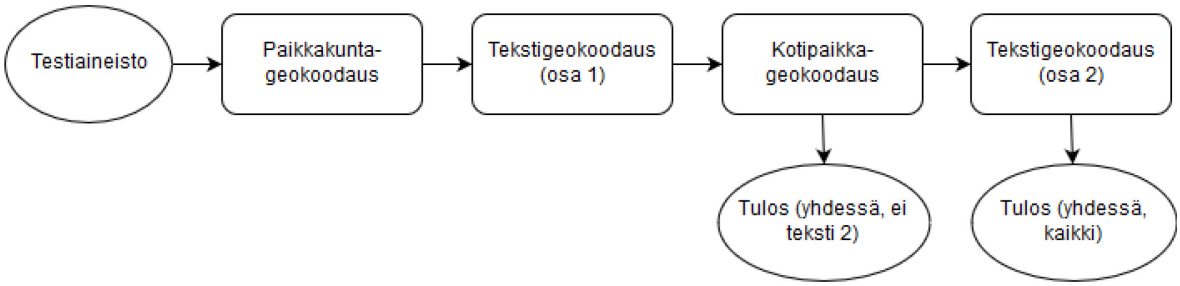
Tässä vaiheessa ei ole toteutettu maan tarkkuudella paikannettujen twiittien esittämistä kartalla. Se vaatii toisen vertailuaineiston käyttöä nykyisen kaupunkiaineiston lisäksi, jotta saadaan maille karttakohteet. Kaupunkiin paikannettu twiitti on luonnollista esittää kartalla pisteenä, mutta maan kohdalla on pohdittava, voidaanko kaikki twiitit sijoittaa maan keskipisteeseen vai pitäisikö ne esittää aluemaisina kohteina. Se onko maan tarkkuudella paikannettu twiitti ylipäätään tarpeen esittää kartalla, riippuu twiittien paikantamisen käyttötarkoituksesta. Jos halutaan esimerkiksi kartoittaa missä päin Suomea Perussuomalaiset herättävät keskustelua ei tieto siitä, että niitä koskeva twiitti on lähetetty Suomesta ole relevantti. Toisaalta jos halutaan kartoittaa minne päin maailmaa suomalaiset matkustavat, niin silloin tulosten ei välttämättä tarvitse olla kaupunkikohtaisia vaan esim. ”Thaimaa” on tarpeeksi tarkka vastaus ja se halutaan siksi esittää kartalla. Yksi mahdollinen tapa esittää maan tarkkuudella paikannetut twiitit on koropleettikartta, jossa väri kertoo, kuinka paljon twiittejä mihinkin maahan on paikannettu.

3.3 Testit

Toteutettuja työkaluja testattiin kappaleessa 3.1 kuvatulla aineistolla, jotta voitaisiin arvioida niiden toimivuutta sekä yhdistämisalgoritmien mahdollisia eroja. Tutkimuksessa testattiin eri kenttiin perustuvia geokoodaustyökaluja erikseen, eli paljonko osumia saadaan, esim. kotipaikkageokoodauksella olettaen, ettei yhdellekään twiitille ole aikaisemmin löytynyt sijaintia (kuva 7). Lisäksi testattiin, minkälaisiin tuloksiin päästään kun käytetään kaikkia geokoodaustyökaluja peräkkäin, kuten niitä on tarkoitus todellisuudessa käyttää (kuva 8). Molemmat variaatiot testattiin erikseen kaikille kolmelle yhdistämisalgoritmille. Testikaavioiden tulosten tarkennukset viittaavat liitteen 1 tulostaulukoihin. Koska paikka- ja kotipaikkakenttien perusteella tehtävät geokoodaukset eivät käytä epätarkkaa vertailua, ovat niiden tulokset luonnollisesti samat lähtökohtien ollessa samat. Erot tekstin tarkan vertailun tuloksissa johtuvat siitä, että jos yksi sana tekstistä on osattu paikantaa tarkasti, mutta jokin toinen myöhempi sana onnistutaan paikantamaan epätarkasti, niin silloin paikannuksen tasoksi tulee 45, eli epätarkka vertailu ja useita vaihtoehtoja. Tämä olisi voitu välttää jättämällä taso tässä tapauksessa päivittämättä, mutta toisaalta käyttäjän on hyvä huomata, että tälle twiitille on olemassa vaihtoehtoisia epätarkempia sijainteja, eikä näin ollen esittää sitä kartalla varmempana sijaintina. Työkalujen yhteiskäytöstä saatujen tulosten osalta tarkastellaan, kuinka suuri osa geokoodatuista twiiteistä sai minkäkin tasoisen tuloksen.



Kuva 7. Työkalujen testaus erikseen



Kuva 8. Työkalujen testaus yhdessä

Tulosten arvioinnissa käytetään mittareina kappaleessa 2.3 kuvailtua täydellisyyttä ja tarkkuutta. Täydellisyys, eli osumaosuus lasketaan sekä niistä twiiteistä jotka on teoriassa mahdollista paikantaa että kaikista twiiteistä, esimerkiksi jos 50 twiittiä sisältää paikkamerkinnän, niin osumaosuus kuvaa, kuinka monta näistä onnistuttiin paikantamaan. Toinen tarkasteltava mittari on tarkkuus. Tarkkuudella tarkoitetaan tässä tapauksessa sitä, kuinka monta paikannetuista twiiteistä on paikannettu oikein. Oikeana ratkaisuna pidetään sitä, miten työn tekijä on twiitit manuaalisesti paikantanut. Luonnollisesti näissä oikeissa ratkaisuissa voi olla inhimillisiä virheitä, joten tarkkuudet pyöristetään kokonaisiin prosentteihin. Lisäksi tarkastellaan virheellisten osumien (false positive) ja puutteiden (false negative) määriä. Virheelliseksi osumaksi lasketaan twiitti, joka on paikannettu väärään paikkaan tai paikannettu vaikka sitä ei olisi pitänyt paikantaa ollenkaan. Oikeiksi osumiksi lasketaan myös sellaiset twiitit, jotka on paikannettu väärään paikkaan, mutta joiden vaihtoehtoisista paikoista löytyy oikea paikka. Puutteita ovat twiitit, jotka olisi pitänyt paikantaa, mutta jolle sitä ei onnistuttu tekemään. Mukaan ei tässä lasketa väärin paikannettuja twiittejä. Puutteiden rajaaminen on haastavaa, sillä riippuu sekä vertailuaineistosta että vertailualgoritmeista, mitkä twiitit voidaan paikantaa. Tässä kuitenkin lasketaan sellaisetkin twiitit, joissa mainitaan paikka, jota ei vertailuaineistossa ole. Koska tutkimusaineisto on kerätty hiihtoloma-aiheisista twiiteistä, mainitaan niissä esimerkiksi hiihtokeskuksia, joita ei kaupunkeja sisältävässä vertailuaineistossa ole. Puutteiden laskemiseksi käytetään kohdassa 3.1 mainittuja itse lisättyjä paikannimiä.

4 Tulosten tarkastelu

4.1 Paikannus kotipaikan perusteella

Taulukot 5 ja 6 esittävät muiden kuin tekstin geokoodauksen tuottamia tuloksia molempien testiaineistojen osalta. Tuloksista voidaan havaita, että sekä paikan että twittertilin kotipaikan perusteella tehdyt geokoodaukset antavat näillä aineistoilla varsin luotettavia tuloksia tarkkuuden ollessa 95-100 % paikannetuista. Se paljonko twiittejä näiden kenttien perusteella pystytään paikantamaan, riippuu täysin aineistosta. Hiihtoloma-aineistossa 74 % twiiteistä sisälsi kotipaikan, kun vastaava luku pakkoruotsi-aineistolle oli 67 %. Ero siinä paljonko twiittejä lopulta pystyttiin paikantamaan kotipaikan perusteella on kuitenkin huomattavasti isompi, kuten taulukoista voidaan nähdä. Yksi mahdollinen syy tälle erolle on käyttäjäryhmä, joka eri aiheisia twiittejä on lähettänyt. Hiihtoloma-aiheisia twiittejä ovat lähettäneet pääasiassa yksityiset henkilöt, jotka esiintyvät omalla nimellään ja usein ilmoittavat kotipaikkansa kaupungin tarkkuudella. Pakkoruotsitwiittejä puolestaan ovat lähettäneet yksityisten henkilöiden lisäksi jonkin ryhmittymän nimellä esiintyvät käyttäjät kuten ”PS-Nuoret” ja käyttäjät kuten ”Ruotsipakko” ja ”Demokraatti”, joiden takana voi olla yksi tai useampi henkilö. Tällaiset käyttäjät näyttävät useimmiten ilmoittavan kotipaikakseen Suomen, sen enempää tarkentamatta.

Taulukko 5. Hiihtoloma-aineiston paikan ja kotipaikan geokoodauksen tulokset

Geokoodausperuste	Osumaosuus kaikista %	Tarkkuus %	Virheelliset esiintymät %	Puutteet %
Paikka	12	100	0	1
Kotipaikka	49	95	5	5
Kotimaa	87	98	2	4

Taulukko 6. Pakkoruotsi-aineiston paikan ja kotipaikan geokoodauksen tulokset

Geokoodausperuste	Osumaosuus kaikista %	Tarkkuus %	Virheelliset esiintymät %	Puutteet %
Paikka	3	100	0	0
Kotipaikka	25	100	0	1
Kotimaa	62	100	0	1

Geokoodauksen ohessa selvitettävä kotimaa vaikuttaa tulosten perusteella olevan luotettavaa tietoa, silloin kun se pystytään selvittämään. On kuitenkin otettava huomioon vertailuaineiston olevan maannimien osalta puutteellinen, mistä johtuen tulos saattaa muuttua jos vertailuaineistoon lisätään enemmän vaihtoehtoisia nimiä tai lyhenteitä maille. Tällöin voi olla syytä lisätä mahdollisuus useammille ehdotuksille kotimaan osalta.

Puutteille tekstipaikannuksen tuloksissa eli sille, ettei työkalu pystynyt paikantamaan twiittiä vaikka siinä manuaalisesti onnistuttiin, on olemassa useita eri syitä. Jos kotimaaksi on paikannettu Suomi, ei muissa maissa olevia paikkoja tarkisteta, esimerkiksi Haaparanta jäi paikantamatta koska se sijaitsee Ruotsissa. Kotimaan virheellinen paikantaminen aiheuttaa vastaavanlaisen tilanteen. Muutamille twittertileille kotipaikka oli ilmoitettu muodossa ”kaupunki, Fin”. Koska vertailuaineistossa ei ole maiden lyhenteitä mukana ei järjestelmä tunnista lyhenteen ”Fin” tarkoittavan Suomea. Sen sijaan Papua-Uusi-Guineassa on Fin niminen kaupunki, joten se asetetaan tilin kotipaikaksi. Tästä johtuen kyseiselle twiitille ei saada tekstigeokoodauksella osumaa, vaikka tekstissä olisi mainittu suomalainen paikkakunta.

Toinen yleinen virhetyyppi on paikat joita ei löydy vertailuaineistosta. Tällaisia ovat esimerkiksi alle 1000 asukkaan paikkakunnat tai lomakeskukset kuten Ylläs. Paikka voi myös olla mainittuna osana pidempää merkkijonoa kuten ”helsinkiairport”, jolloin epätarkka vertailualgoritmin antama samanlaisuusarvo ei ylitä käytettävää raja-arvoa. Koska päädyimme suoritusajan lyhentämiseksi vertailemaan sanoja vain niihin kaupunginnimiin jotka alkavat samalla kirjaimella, ei vertailua Helsinkiin päästä edes tekemään jos tekstissä esiintyy hashtag ”visithelsinki”.

Geokoodauksessa oleellisesti tuloksiin vaikuttava tekijä on mihin vedetään osuman raja epätarkkaa vertailua tehtäessä. Jos raja on löysä, löydetään aineistosta varmemmin kaikki paikkakunnat, mutta mukaan tulee huomattava määrä vääriä osumia. Jos puolestaan raja on tiukka saattaa osa tekstin paikkakunnista jäädä löytämättä, mutta toisaalta myös väärien osumien määrä on pienempi, mistä seuraa tulosten tarkkuusprosentin paraneminen.

4.2 Yhdistämisalgoritmien vertailu

Todennäköisyyspohjaisten yhdistämisalgoritmien Levenshteinin etäisyys, LCS ja n-grammi vertailu suoritettiin hiihtoloma-aineistoa käyttämällä. Tarkka vertailu tapahtuu kaikilla menetelmillä samalla tavalla ja se tuotti noin 70 % tarkkuudella oikeita vastauksia. Vaihtelut johtuvat aikaisemmin kuvatusta geokoodaustason päivittämisestä. Jos tarkastellaan pelkästään epätarkalla vertailulla saatuja tuloksia ovat erot sen sijaan huomattavan paljon suuremmat. Taulukko 7 esittää epätarkan vertailun tulokset menetelmittäin sekä tarkan ja epätarkan vertailun yhteenlasketun tarkkuuden.

Taulukko 7. Epätarkan vertailun tulokset menetelmittäin

Menetelmä	Osumaosuus, %	Tarkkuus, %	Oikeat, kpl	Yhteenlaskettu tarkkuus, %
Levenshtein 0.25	7	15	9	57
Levenshtein 0.17	3	22	5	65
LCS	17	22	34	46
Digram	4	34	11	65
Trigram	1	83	5	71

Osuman rajana käytetyllä arvolla on keskeinen merkitys sekä tulosten määrässä että laadussa. Levenshteinin yhdistämisalgoritmia kokeiltiin kahdella eri raja-arvolla. Käytettyihin raja-arvoihin kaikkien työkalujen osalta päädyttiin työkalujen kehityksen aikana tehtyjen testien perusteella. Optimaalisen raja-arvon löytäminen on haastavaa, sillä rajaa löysäämällä voi saada mukaan muutaman osuman lisää, mutta samalla tulee huomattava määrä virheellisiä osumia. Yhtä oikeaa raja-arvoa kullekin menetelmälle tuskin on olemassa, vaan se riippuu aineistosta. Tämän työn testiaineistot olivat pääasiassa suomeksi, mutta jollain toisella kielellä kirjoitettujen twiittien geokoodaukseen raja-arvoksi todennäköisesti sopii joku muu.

Ensimmäisessä Levenshteinillä tehdyssä testissä osuman raja oli 0,25, jos kotimaa on tiedossa (eli ensimmäisen vertailun kohdalla) ja 0,17 toisella yrityksellä. Toisessa testissä raja-arvo oli molemmissa tapauksissa 0,17. Korkeampi raja-arvo tuotti, kuten olettaa sopii, enemmän osumia. Oikeiden osumien määrä ei kuitenkaan ole merkittävästi korkeampi kuin alemmalla raja-arvolla. Oikein paikannettujen twiittien ero, kun tarkat ja epätarkat lasketaan yhteen, on vain kaksi kappaletta enemmän korkeammalla raja-arvolla. Samalla virheellisten

osumien määrä on korkeammalla raja-arvolla isompi, joten tarkkuus on heikompi. Näin ollen ainakin tällä aineistolla päästään parempaan tarkkuuteen käyttämällä alhaisempaa raja-arvoa, ilman että se merkittävästi lisää puutteiden määrää.

N-grammi vertailulla kokeiltiin sekä digrammi että trigrammi vertailua. Molemmille käytettiin raja-arvoa 0,4. Trigrammi tuotti vain kuusi osumaa vastaan digrammin 32 kpl, mutta trigrammin osumat olivat yhtä lukuun ottamatta kaikki oikein. Tarkkuus oli täten trigrammilla 83 %, mikä on huomattavasti kaikkia muita menetelmiä enemmän. Puutteiden määrä oli kuitenkin trigrammilla 6 kpl enemmän, mikä tarkoittaa, että digrammi osasi yhdistää 6 paikannimeä, joita trigrammi ei osannut yhdistää. Kaikista twiiteistä ero puutteiden määrässä on kuitenkin vai noin 1 %.

LCS osoittautui selkeästi parhaaksi jos halutaan löytää mahdollisimman paljon oikeita osumia. LCS toimii paremmin kuin muut etenkin lyhyiden paikannimien yhdistämisessä, esimerkiksi ”Inarissa” onnistuttiin yhdistämään paikkaan ”Inari”, vaikka taivutus on melko iso osa sanaa ja mm. digrammi ei näitä osannut yhdistää. Sama pätee aineistossa esiintyneeseen hashtagiin ”lahtimm”, jonka LCS onnistui paikantamaan Lahteen, toisaalta ”mmlahti” puolestaan paikantui virheellisesti Maalahteen. Osumien määrän lisääntyessä virheellisten osumien määrä kasvaa muihin menetelmiin verrattuna. Tarkkuusprosentti on kutenkin samaa luokkaa Levenshtein 0,17 kanssa. Tekstipaikannuksen yhteenlaskettu tarkkuus (46 %) on kuitenkin selkeästi tämän tutkimuksen huonoin. LCS:n käyttö vaatii siis huomattavan määrän tulosten manuaalista perkaamista, jos halutaan esittää vain oikein paikannetut twiitit kartalla. Oikeiden vastausten huomattavasti muita isompi määrä, selittyy osittain sillä, että tarkan osuman lisäksi samasta twiitistä on myös paikannettu jokin muu sana epätarkalla vertailulla. LCS:n raja-arvona käytettiin ensimmäisen tekstivertailun osalta 0,25 ja toisen 0,15.

Parhaaseen tarkkuuteen päästään trigrammilla, joskin ero kun tarkastellaan koko tekstipaikannuksen tuloksia ei ole yhtä suuri kuin pelkästään epätarkkaa vertailua tarkasteltaessa. Toiseksi parhaita, miltei samoilla yhteenlasketuilla prosenttiluvuilla, ovat Levenshtein (0,17 raja-arvolla) ja digrammi. Jos tarkastellaan vain epätarkkaa vertailua, on digrammi kuitenkin selkeästi toiseksi paras vaihtoehto, sen tuottaessa toiseksi eniten oikeita vastauksia, samalla kun sen tarkkuusprosentti on toiseksi paras. Koska epätarkan vertailun tuottamien oikeiden paikannusten määrä liki tuhannen twiitin vertailuaineistossa on vain 5-34 kpl, ei tästä voida kuitenkaan kovin pitkälle meneviä johtopäätöksiä tehdä, sillä yksittäisten twiittien aiheuttamat vaihtelut tuloksissa ovat niin suuret. Voidaan kuitenkin todeta, että huolimatta aineiston sisältämästä poikkeuksellisen suuresta määrästä paikkamainintoja, ovat ne jotka pystytään epätarkalla yhdistämisalgoritmilla tunnistamaan kuitenkin aika vähissä. Perusmuodossa olevat paikannimet ovat twiiteissä, mahdollisesti jopa yleisempiä kuin tavallisessa tekstissä sillä usein paikka mainitaan twiitissä hastagilla, esim. #Tohmajärvi tai #Tampere, jolloin paikannimi voidaan aineiston normalisoinnin ansiosta yhdistää vertailuaineistoon tarkalla vertailulla.

Kolmea hiihtoloma-aineistolla parhaiten toiminutta menetelmää, eli Levenshtein (0,17), digram ja trigram, kokeiltiin pakkoruotsiaineistolle. Tulosten ollessa hiihtoloma-aineistolla lupaavia, ovat ne pakkoruotsi-aineistolla selvästi heikompia. Aineiston reilusta tuhannesta twiitistä vain 10 sisältää manuaalisesti tunnistetun paikkailmaisun, mikä tarkoittaa, ettei tekstipaikannuksella tulisi saada kovinkaan paljon osumia. Tekstigeokoodaustyökalut tunnistavat paikka-ilmaisuista vain 1-3 kpl menetelmästä riippuen. Siinä paljonko vääriä osu-

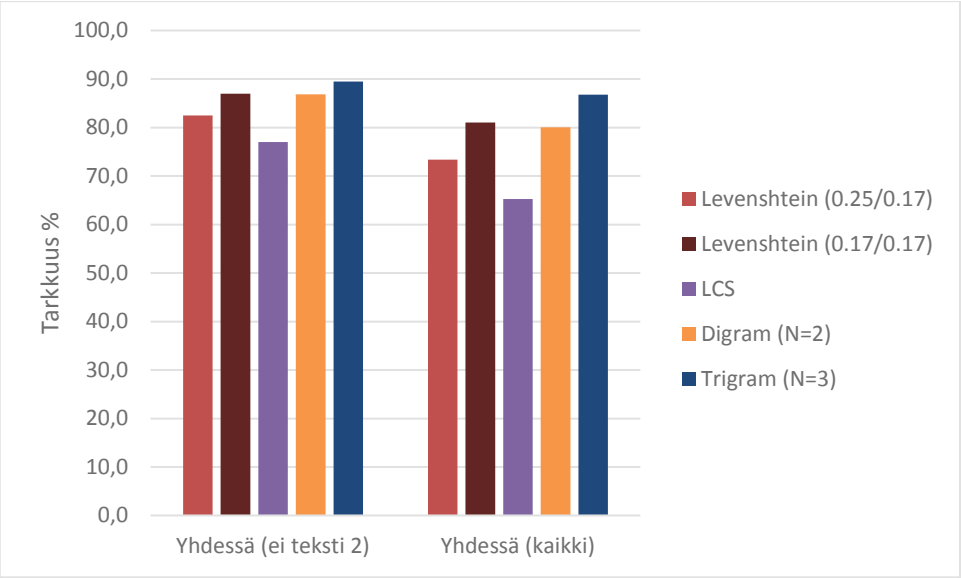
mia eri menetelmät tuottivat, on sen sijaan suurempaa vaihtelua. Digrammi paikansi epätarkalla vertailulla 60 twiittiä, joista 2 oikein, kun taas trigrammi paikansi vain yhden ja se oli väärin. Levenshtein puolestaan paikansi 25 twiittiä, joista yksi oli oikein, ja sijoittui näin ollen edellä mainittujen väliin.

Huomattavinta pakkoruotsi-aineiston tekstigeokoodauksen tuloksissa on kuitenkin tarkan vertailun heikko tulos. Kun hiihtoloma-aineiston kohdalla tarkkuus oli 70 % luokkaa, on tarkan vertailun tarkkuus pakkoruotsi-aineistolle ainoastaan vajaan prosentin. Tarkalla vertailulla löytyi 105 osumaa, joista vain yksi oli oikein. Tämä tarkoittaa hukkasanalistan käytön tavallisten sanojen rajoittamiseen olevan riittämätön toimenpide, jonka vuoksi siihen on keksittävä myös muita keinoja. Virheellisesti paikannettujen erillisten sanojen määrä ei kuitenkaan ole valtava, sillä twiiteissä toistuvat sanat kuten: musta. lista, paras, lue ja osa, joista johtuen lukuisat twiitit paikannetaan virheellisesti samaan paikkaan. Jos paikannettavien sanojen pituuden alarajaa nostettaisiin kolmesta kirjaimesta viiteen, niin virheellisesti paikannetut sanat vähenisivät vielä jonkin verran, sillä mitä lyhyempi sana sitä todennäköisemmin, jostain löytyy samanniminen paikka. Kuten todettua se kuitenkin vähentää myös oikeiden osumien määrää. Epätarkalla vertailulla virheellisesti paikannettujen sanojen joukossa esiintyy myös pidempiä sanoja, eikä samankaltaista korrelaatiota ole havaittavissa.

4.3 Kokonaisuuksien vertailu

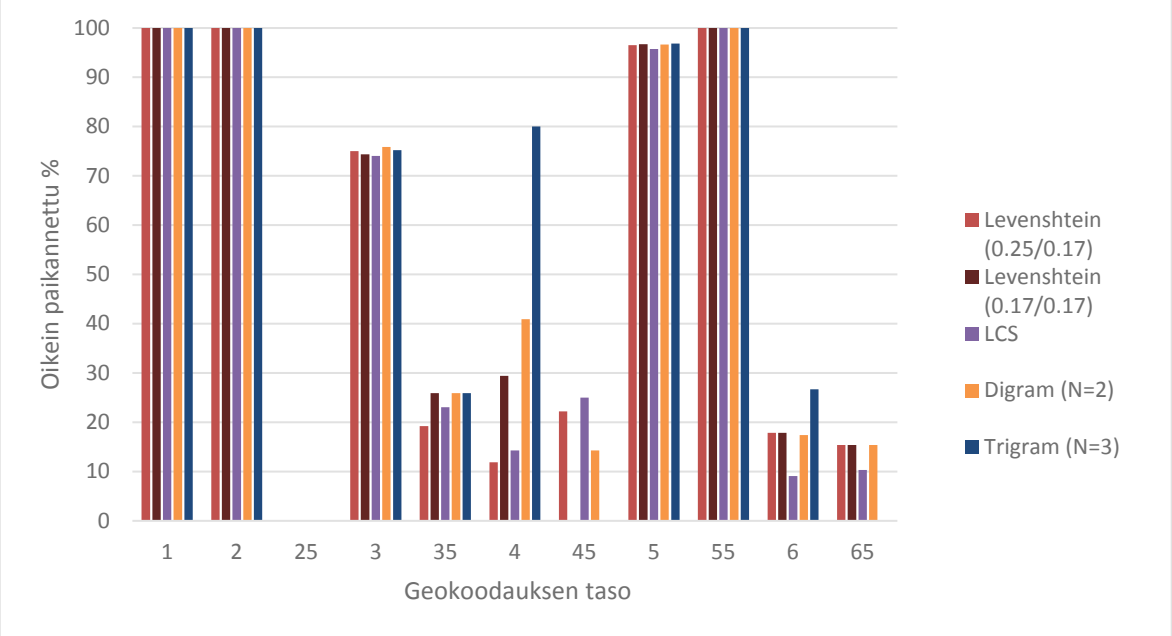
Kehitettyjen geokoodaustyökalujen varsinainen tarkoitus on saada aikaan mahdollisimman hyvä tulos käyttämällä useita työkaluja samalle aineistolle, jolloin ratkaisevaa ei ole ainoastaan yhden työkalun toimivuus. Ideaalitilanteessa pystytään paikantamaan kaikki twiitit, joille käyttäjä on ilmoittanut paikan. Tämän jälkeen paikannetaan lopuista twiiteistä kaikki ne, joiden tekstissä mainitaan paikka. Lopuksi paikannetaan vielä paikantamatta olevat twiitit kotipaikan perusteella. Koska seuraava paikannus tehdään aina niille twiiteille joita ei ole vielä paikannettu, mahdollisimman harvojen twiittien virheellinen paikantaminen on tärkeää.

Kuva 9 esittää geokoodaustyökalujen yhteiskäytön tulokset hiihtoloma-aineistolle. Huomautus ”ei teksti 2” tarkoittaa, että jälkimmäistä tekstinpaikannusta ei ole suoritettu. Tarkimmaksi osoittautuu testi, jossa tekstivertailuun on käytetty trigrammia. Yksi syy tähän saattaa olla nimenomaan, että trigrammi tuottaa vähiten virheellisiä paikannuksia ja näin ollen isompi osa twiiteistä paikannetaan kotipaikan perusteella, jolloin tarkkuus on kuten todettua parempi. Samasta syystä LCS puolestaan on selkeästi huonoin.



Kuva 9. Hiihtoloma-aineiston geokoodauksen tulokset kaikkia työkaluja käytettäessä.

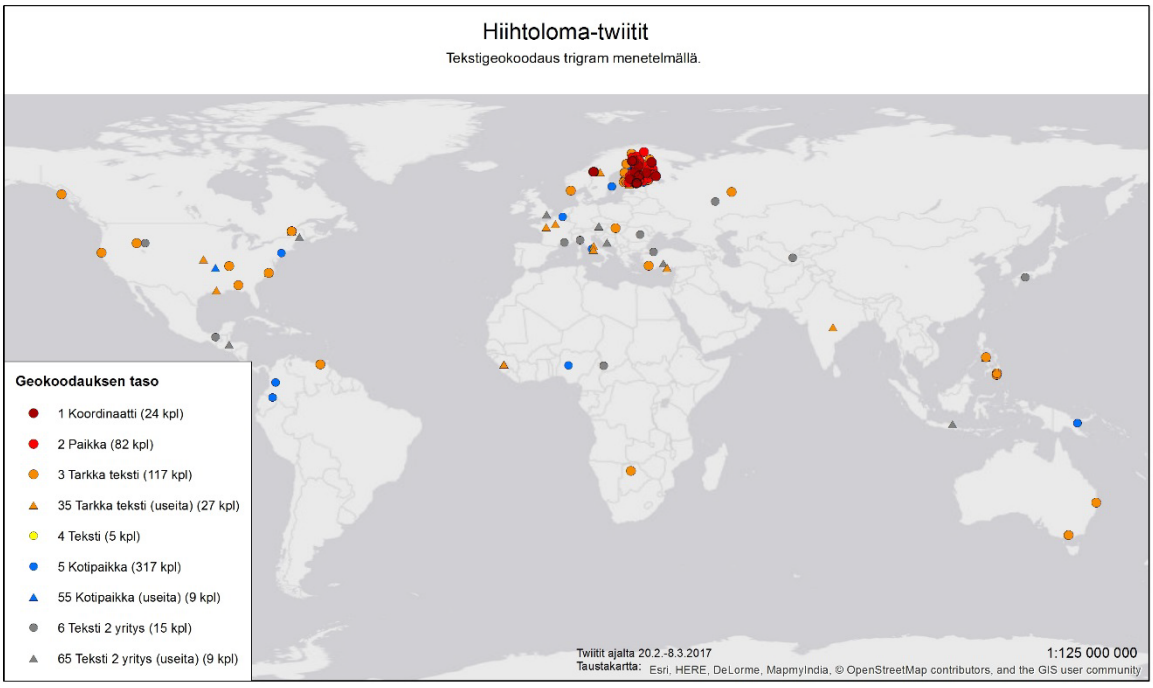
Kuva 10 näyttä kunkin menetelmän tarkkuuden jaoteltuna paikannuksen tasoihin. Siitä voidaan nähdä, että kaikki koordinaattien tai paikan perusteella geokoodatut sijainnit ovat oikein, kun taas epätarkat vertailut (tasot 4, 45, 6 ja 65) aiheuttavat enemmän hajontaa. Tason 25 nolla prosenttia johtuu siitä, ettei aineistossa esiintynyt paikkakentässä yhtään paikannimeä, joka täsmää useampaan vertailuaineiston paikkaan. Sama pätee muutamiin muihin nollan prosentin tuloksiin.



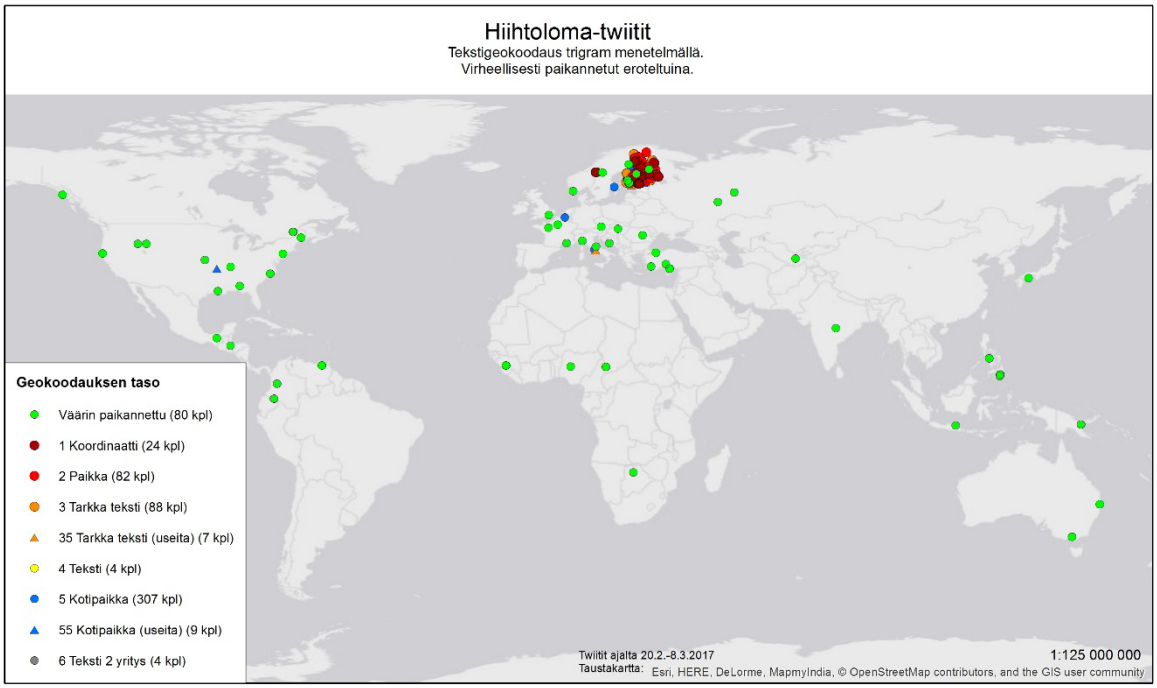
Kuva 10. Hiihtoloma-aineiston yhteisgeokoodauksen tasokohtaiset tarkkuudet.

Kuvissa 11 ja 12 nähdään trigrammilla tehdyn geokoodauksen tulos kartalla. Kuvan 11 kartalla twiitit näytetään niiden saaman paikannustason mukaan. Kuva 12 on saman geokoodauksen tulos, joka on käyty manuaalisesti läpi ja johon on merkattu kaikki virheellisesti paikannetut twiitit vihreällä. Tästä kuvasta voidaan havaita, että muutamia poikkeuksia lukuun ottamatta kaikki väärin paikannetut twiitit sijaitsevat jakautuneena ympäri maailmaa,

kun taas oikein paikannetut ovat pääosin keskittyneenä Suomeen. Karttoja vertailemalla sekä liitettä 3 tutkimalla nähdään valtaosan väärin paikannetuista twiiteistä olevan seurausta tarkasta vertailusta. Kuten todettua tarkka vertailu, joka suoritetaan koko vertailuaineistolle johtaa useiden tavallisten sanojen virheelliseen paikantamiseen. Muiden menetelmien kohdalla, joissa epätarkan vertailun tuottamien osumien määrä on trigrammia suurempi, tuottaa myös epätarkka vertailu huomattavan määrän virheellisiä tuloksia. Tarkan vertailun osalta on havaittavissa, että jos samalle twiitille on löytynyt useita vaihtoehtoisia tarkkoja osumia niin todennäköisyys, että joku niistä on oikein on noin 23-26 %. Niiden twiittien osalta joille puolestaan on saatu vain yksi tarkka osuma, on tarkkuus sen sijaan 74-75 %. Twiitit joille löytyi useampi tarkka osuma ovat pääsääntöisesti niitä joiden kotimaa ei ole tiedossa ja joita sen vuoksi verrataan koko vertailuaineistoon. Näin ollen niissä saattaa olla useita tavallisia sanoja jotka paikannetaan. Tällaisia ovat esimerkiksi ”mut”, ”sit” ja ”ens”. Toisaalta useat tavalliset suomenkielen sanat esiintyvät paikanniminä useammassa kuin yhdessä paikassa, esimerkiksi ”loma” tuottaa seitsemän tarkkaa osumaa ja ”ale” kaksi. Myös epätarkan vertailun osalta on havaittavissa samankaltaista trendiä kaikkien muiden menetelmien paitsi LCS:n osalta. Ero yhden ja usean epätarkan osuman tarkkuusprosenteissa ei kuitenkaan näillä ole yhtä merkittävä.

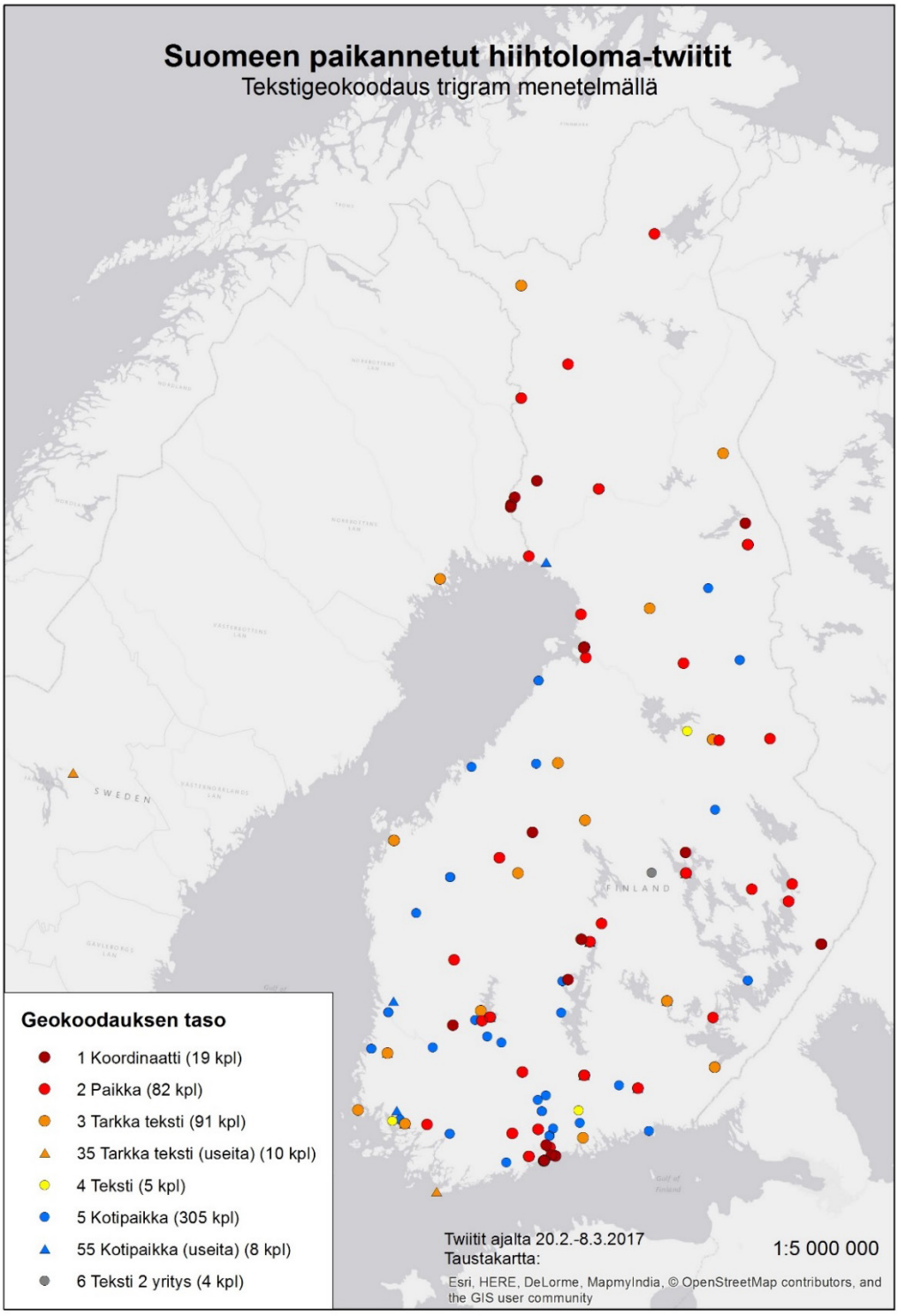


Kuva 11. Hiihtolomatwiittien trigrammigeokoodauksen tulos. Isompi kuva liitteessä.



Kuva 12. Hiihtolomatwiittien trigrammigeokoodauksen tulos. Väärin paikannetut vihreällä.

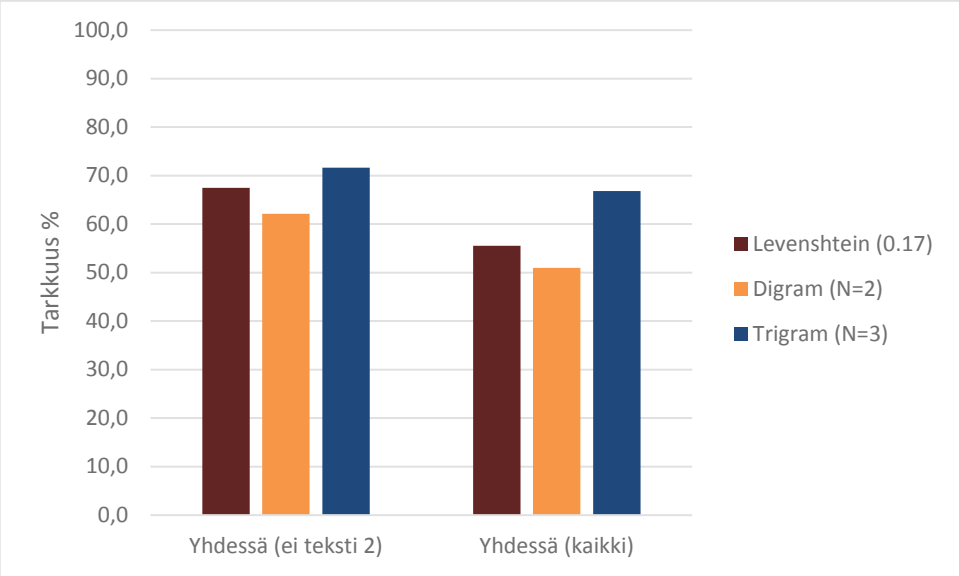
Tuloksen visualisoinnille tuottaa haastetta geokoodattujen pisteiden hajanainen ja epätasainen jakautuminen eri puolille maailmaa. Koska testiaineisto koostuu pääasiassa suomenkielisistä twiiteistä sijoittuvat useimmat pisteet Suomeen, mutta etenkin virheellisesti geokoodatut twiitit hajautuvat laajalle alueelle. Tuloksia on tästä syystä helpointa tarkastella Arcmapissa, jolloin käyttäjä pystyy lähentämään karttakuvaa haluamiinsa paikkoihin sekä tarkastelemaan twiittien sisältöä ja metatietoja. Staattisten karttojen tarkoituksena on tässä tapauksessa lähinnä antaa yleiskuva twiittien sijoittumisesta, eikä mahdollistaa yksittäisen twiitin tarkan paikan tarkastelua. Kuva 13 esittää samaa tulosta kuin kuva 11, mutta vain Suomen osalta, jolloin on mahdollista erottaa twiittien jakautuminen eri puolille maata. Kartalla esitettynä samaan paikkaan geokoodatuista twiiteistä näkyy vain yhden twiitin paikannustaso. Tämä saattaa hämätä luulemaan, että esimerkiksi kaikki Ouluun paikannetut twiitit on paikannettu annettujen koordinaattien perusteella, vaikka joukossa on myös epätarkalla vertailulla sekä kotipaikan perusteella paikannettuja twiittejä.



Kuva 13. Trigrammilla Suomeen geokoodatut hiihtoloma-twiitit.

Tutkimuksessa kokeiltiin kolmea parhaiten hiihtoloma-aineistolle toiminutta menetelmää pakkoruotsiaineistolla ja näiden testien tulokset näkyvät kuvassa 14. Kuten tekstigeokoodauksen tuloksista voidaan päätellä, ovat tulokset tällä aineistolla hiihtoloma-aineistoa heikommat. Kotipaikkageokoodauksen tarkat tulokset kuitenkin pelastavat yhteisgeokoodauksessa paljon ja näin ollen testattujen menetelmien tarkkuudet vaihtelevat kaikkia työkaluja käytettäessä 51 ja 67 prosentin välillä. Kun tarkastellaan yhteistulosten tarkkuutta, on kuitenkin otettava huomioon, että oikeita tuloksia tarkasteltaessa on käyty läpi geokoodauksen

tasot yksitellen. Tämä tarkoittaa, että tasolla 5, eli kotipaikan perusteella geokoodattu twiitti voi olla oikein, mutta oikeastaan se olisi pitänyt onnistua paikantamaan jo edellisessä vaiheessa, eli epätarkalla vertailulla. Tällaisten twiittien määrä riippuu siitä, paljonko paikannimiä tekstissä esiintyy. Hiihtoloma-aineistossa näiden twiittien lukumäärä liikkuu noin 50 kappaleessa menetelmästä riippuen ja pakkoruotsiaineiston kohdalla niitä on puolestaan vain 5-6.



Kuva 14. Pakkoruotsi-aineiston geokoodauksen tulokset kaikkia työkaluja käytettäessä.

4.4 Suoritusajat

Testien suoritusaikoja tarkasteltiin, jotta voidaan nähdä mihin vaiheisiin aikaa kuluu minkäkin verran ja onko eri menetelmien välillä mahdollisesti eroja, vaikka suoritusajan optimoiminen ei ollutkaan työn painopisteenä. Testit suoritettiin normaalilla työasemalla, jossa ei samaan aikaan ollut käynnissä muita raskaita prosesseja ja ajat ovat ArcMapin ilmoittamia suoritusaikoja.

Vaajaan tuhannen twiittin normalisointiin aikaa kului vajaa minuutti. Paikka-kentän perusteella tehtävän geokoodauksen vaatima aika riippuu pitkälti siitä, monessako twiitissä on paikka mainittuna. Tehdyissä testeissä kesto vaihteli 15-30 sekunnin välillä. Kotipaikan perusteella tehtävän geokoodauksen suoritusaika puolestaan oli testiaineistoilla 1-2 minuuttia. Kuten on käynyt moneen otteeseen ilmi, on tekstigeokoodaus selkeästi eniten aikaa vievä prosessi ja siksi sitä onkin yritetty monin eri keinoin nopeuttaa. Hiihtoloma-aineostolle puhtaalta pöydältä tehdyn tekstigeokoodauksen kesto on esitetty taulukossa 8. Siitä nähdään, että ensimmäisen tekstigeokoodaustyökalun osalta lyhyimmän ja pisimmän suoritusajan välinen ero on vain 55 sekuntia. Twiittien keskimääräinen käsittelyvauhti ensimmäisessä vaiheessa oli 1,8 sekuntia/twiitti. Toisen työkalun osalta ero eri menetelmien välillä on hieman yli 10 minuuttia, mikä on jo huomattava. N-grammien vertailu vaikuttaa olevan huomattavasti nopeampaa kuin sanojen vertailu Levenshtein menetelmällä. LCS puolestaan sijoittuu ajankäytössä näiden välille. Samankaltaisia tuloksia antavat testit pakkoruotsi-aineistolla, joskin siinä ero digramin ja trigrammin välillä on useita minuutteja digrammin hyväksi. Ensimmäisen vaiheen vähäisiä eroja saattaa selittää se, että siinä tehdään epätarkan vertailun lisäksi useita muita vaiheita, jotka toteutetaan samalla tavalla jokaisella työkalulla. Ensiksi

selvitetään twiittien kotimaa, jonka jälkeen suoritetaan tarkka vertailu ja vasta lopuksi epätarkka vertailu niille twiiteille, joille kotimaa löytyi.

Taulukko 8. Tekstigeokoodaustyökalujen suoritusajat hiihtoloma-aineistolle.

Vertailualgoritmi	Tekstigeokoodaus vaihe 1	Tekstigeokoodaus vaihe 2
Levenshtein (0,25)	27 min 32 sek	23 min 28 sek
Levenshtein (0,17)	28 min 10 sek	23 min 43 sek
LCS	27 min 15 sek	17 min 24 sek
Digram	27 min 21 sek	13 min 14 sek
Trigram	27 min 37 sek	13 min 29 sek

Kun kaikki työkalut suoritetaan peräkkäin, myöhemmin suoritettavan työkalun suoritusaika luonnollisesti lyhenee, koska sen tarvitsee käydä läpi vain ne twiitit, joita ei ole vielä paikannettu. Koska yhteenlaskettu suoritusaika kaikille työkaluille koostuu pääasiassa tekstigeokoodauksista, seuraa se samaa kuviota kuin edellä. Kokonaisaika vaihtelee digrammin 38 minuutista ja 50 sekunnista Levenshteinin (0,17) 48 minuuttiin ja 55 sekuntiin. Koko prosessin suorittamiseen menee kuitenkin luonnollisesti enemmän aikaa, koska työkalut on käynnistettävä ja tuloksille on määriteltävä tiedostonimet jne. Merkittävin lasku suoritusajassa saatiin kehityksen aikana aikaan lisäämällä ehto, jonka mukaan epätarkka vertailu suoritetaan ainoastaan niiden nimien kanssa jotka alkavat samalla kirjaimella kuin tarkasteltava sana sekä rajoittamalla vertailtavien sanojen pituuseroa. Tulokset ovat samansuuntaisia kuin aikaisemmin tehdyissä tutkimuksissa, sillä myös Ranzijn (2013) toteaa LCS:n olevan hitaampi kuin trigrammi.

5 Ehdotuksia tulevaisuutta varten

Geokoodaustyökalujen tuottamien tulosten parantamiseksi on tulevaisuudessa pystyttävä erottamaan tekstistä paikkaan viittaavat sanat ja geokoodattava vain ne, jotta virheellisesti paikannettujen tavallisten sanojen määrä saadaan alas. Tätä varten on hyödynnettävä jotain kappaleessa 2.5.3 esiteltyä menetelmää. Monikielinen aineisto tuottaa haasteita sekä sääntöpohjaiselle geoparsingille että sanastojen käytölle. Jos päädytään käyttämään jompaa kumpaa näistä, on syytä joko kääntää kaikki twiitit englanniksi, kuten Zhang ja Gelernter (2013) ehdottavat, tai keskittyä kehittämään menetelmä joka toimii muutamilla eri kielillä mutta ei kaikilla. Koneoppimista hyödyntävä geoparsing on selvästi suosituin tällä hetkellä, sillä useimmat viimevuosien tutkimukset ovat käyttäneet siihen perustuvaa menetelmää. Koneoppiminen on varteenotettava vaihtoehto, jos käytettävissä on tarpeeksi laaja opetusaineisto, jota pystytään tarvittaessa päivittämään. Geoparsingmenetelmä, joka osaa erotella tekstistä paikannimet eikä ota mukaan kovin paljon muita sanoja, vähentäisi sekä tarkalla että epätarkalla vertailulla saatavien virheellisten tulosten määrää ja parantaisi näin ollen tulosten tarkkuutta. Toisaalta jos geoparsing ei osaa tunnistaa kaikkia paikkaan viittaavia sanoja vaan jättää osan valitsematta, niin se saattaa lisätä puutteiden määrää tuloksissa verrattuna kaikkien sanojen paikantamiseen.

Twiiteissä mainitaan usein paikkoja, jotka ovat laajuudeltaan kaupungin ja maan välissä, kuten Lappi, Länsi-Suomi ja Uusimaa. Siksi vertailuaineistoa tulisikin jatkossa laajentaa koskemaan esimerkiksi maakuntia. Maakuntien osalta aluejako on melko selkeä. Kun joku kirjoittaa olevansa Pirkanmaalla pystytään alue rajaamaan tarkasti, kunhan vain oikea vertailuaineisto on käytössä. Kuntaliitokset aiheuttavat kuitenkin, paitsi kuntien, myös maakuntien rajoihin muutoksia, jolloin vertailuaineiston ja geokoodattavan syötteen ajallinen sijoittuminen on merkittävä tekijä. Jos halutaan geokoodata 30 vuotta vanha dokumentti, jossa esiintyy kuntien nimiä, ei voida käyttää nykyistä kuntajakoa vertailuaineistona. Epätarkempia ilmaisuja, kuten esimerkiksi Länsi-Suomi, on vielä hankalampi rajata. Ainakaan viimeksi voimassa ollutta läänijakoa ei voida käyttää, sillä siinä Länsi-Suomi ulottuu Jyväskylään saakka ja kukaan Jyväskylässä oleva tuskin kirjoittaisi olevansa Länsi-Suomessa. Tällaiset ilmaisut riippuvat aina käyttäjän näkökulmasta. Pyydettäessä helsinkiläistä ja oululaista piirtämään Pohjois-Suomi kartalle ovat tulokset todennäköisesti aivan erilaisia. Koska eri maissa on lisäksi erilaiset tavat tehdä aluejakoja, on maailmanlaajuisen vertailuaineiston kehittäminen haastavaa. Esimerkiksi Yhdysvalloissa osavaltiot ovat tärkeässä roolissa ja ne mainitaan usein erottamaan samannimisiä kaupunkeja toisistaan.

Geokoodaustyökaluja jatkokehitettäessä on syytä harkita toisenlaista tietorakennetta kuin tässä käytetty dictionary, jotta vertailuaineistosta voidaan hakea vastaavuuksia ilman, että koko aineisto on käytävä läpi. Yksi mahdollisuus on puurakenne, jossa maat ovat ylemmällä tasolla ja kaupungit niiden alla. Välissä voisi olla muitakin paikannuksen tasoja, kuten edellä mainitut maakunnat ja osavaltiot. Puurakenteessa kunkin tason kohteet järjestetään aakkosjärjestykseen, jotta kaikkia ei tarvitse käydä läpi vaan voidaan siirtyä suoraan oikean kirjaimen kohdalle. Haastetta tässä kuitenkin aiheuttavat vaihtoehtoiset nimet, jotka monissa tapauksissa alkavat eri kirjaimella kuin varsinainen nimi, koska ne ovat eri kielellä. Jos esimerkiksi verrataan merkkijonoa ”pietarsaari” kaikkiin Suomessa oleviin p:llä alkaviin kaupunkeihin, niin osumaa ei löydy, vaan se löytyy Jakobstadin vaihtoehtoisia nimiä tarkasteltaessa. Oman lisähaasteensa tuovat paikannimet jotka on kirjoitettu muilla kuin latinalaisilla aakkosilla, esimerkiksi arabiaksi tai kiinaksi. Myös osalle Suomen kaupungeista löytyy Geonamesista vaihtoehtoiset nimet näillä kielillä. Hakemista puurakenteesta voidaan kuvata neljällä eri esimerkkitapauksella.

Tapaus 1: Haettava: Kokkola. Maa: Suomi. Haetaan puusta Suomi ja sen alta k:lla alkavat kaupungit, jotka käydään läpi.
Tapaus 2: Haettava: Kokkola. Maa: ei tiedossa. Käydään läpi jokaisen maan kohdalta kaikki k:lla alkavat kaupungit.
Tapaus 3: Haettava: Pietarsaari. Maa: Suomi. Käydään ensin läpi Suomen p:llä alkavat kaupungit. Koska osumaa ei löytynyt käydään tämän jälkeen läpi kaikkien Suomen kaupunkien vaihtoehtoiset nimet.
Tapaus 4: Haettava: Pietarsaari. Maa: ei tiedossa. Käydään läpi kaikkien maiden p:llä alkavat kaupungit. Koska osumaa ei löytynyt käydään seuraavaksi läpi kaikkien maiden kaupunkien vaihtoehtoiset nimet.

Toinen parannettava asia on moniosaisten paikannimien, kuten New York City, tunnistaminen, sillä niitä ei tässä työssä kehitetyn menetelmän avulla pystytä tekstistä paikantamaan. Tekstin vertailu tapahtuu aina sana kerrallaan ja siksi vertailuaineistoon verrataan ensin ”new”, sitten ”york” ja lopuksi ”city”. Näistä York ja City löytyvät Geonamesista, mutta ne viittaavat muihin kaupunkeihin, kuin siihen mitä haettiin. Tämän ongelman ratkaisemiseksi voitaisiin käyttää saman tyyppistä tokenisointiprosessia kuin mitä Zhang ja Gelernter (2013) käyttävät omassa työssään. Tällöin tekstiä ei jaeta jokaisen välilyönnin kohdalta vaan se pyritään jakamaan siten että yhteen kuuluvat merkkijonot pysyvät yhdessä.

Yhtenä ratkaisuna tarkan vertailun väärin paikantamille sanoille voisi käyttää aineistokohtaista poistosanalistaa. Tällöin tekstigeokoodaus ajettaisiin kerran läpi, jonka jälkeen selattaisiin tulokset läpi, jotta huomataan, esiintyykö niissä toistuvasti samoja väärin paikannettuja sanoja. Nämä sanat lisättäisiin poistolistaan, jonka jälkeen tekstin geokoodaus suoritettaisiin uudestaan alkuperäiselle aineistolle, mutta poistettavat sanat jäisivät pois. Tämä on kuitenkin hidas ja manuaalista työtä vaativa prosessi, mikä vuoksi se ei sovellu lopulliseksi ratkaisuksi.

6 Johtopäätökset

Geokoodaus on prosessi, jossa tekstille (osoitteelle tai paikannimelle) selvitetään koordinaatit. Operaation taustalla on useita eri vaiheita joiden määrä ja monimutkaisuus vaihtelevat paikannettavasta syötteestä riippuen. Geokoodaus aloitetaan normalisoimalla syöte sekä vertailuaineisto. Normalisoinnin tarkoituksena on saattaa aineistot vertailukelpoisiksi poistamalla ylimääräiset erikoismerkit ja yhtenäistämällä lyhenteiden käyttö. Vertailuaineisto koostuu paikannimistä ja niitä vastaavista maantieteellisistä kohteista piste-, viiva- tai aluemuodossa. Normalisoinnin jälkeen jokainen osoite tai paikannimi pyritään yhdistämään vertailuaineiston vastaavaan kohteeseen, deterministisellä- tai todennäköisyyspohjaisella yhdistämismenetelmällä, tai molemmilla. Vertailuaineiston koostuessa viiva tai aluekohteista on yhdistämisen tuloksille vielä suoritettava interpolaatio, jos tulokseksi halutaan yksittäisen pisteen koordinaatit.

Osoitteiden geokoodaus suoritetaan useimmiten katuverkkoon perustuvalla menetelmällä, jolloin lineaarisella interpoloinnilla voidaan saada koordinaatit myös sellaisille osoitteille, jotka eivät oikeasti ole olemassa. Katuverkkomenetelmä tuottaa tarkimmat tulokset tiheään asutuilla alueilla. Aluemenetelmässä osoite kohdistetaan kiinteistöön ja piste interpoloidaan sen keskelle. Osoitepistemenetelmä tuottaa tarkimpia tuloksia, mutta sen käytössä ongelmaksi muodostuu laajoja alueita kattavien aineistojen heikko saatavuus.

Twiittien paikannusta niissä mainitun paikan tai twittertilin kotipaikan perusteella voidaan pitää luotettavana, sillä molempien testiaineistojen osalta tarkkuus oli yli 95 %. Jos twiittiin on liitetty paikka tai tilin kotipaikaksi on ilmoitettu todellinen paikka, voidaan siis twiitti hyvällä todennäköisyydellä paikantaa oikein. Kuinka suuri osuus twiiteistä pystytään näiden tietojen perusteella paikantamaan, riippuu aineistosta, eli käytännössä siitä kuka twiitit on lähettänyt. Omalla nimellä esiintyvät yksityishenkilöt näyttävät useammin mainitsevan kotipaikkakunnan, kuin ryhmät tai peitenimellä esiintyvät tahot.

Kun geokoodattavana on teksti, on ennen yhdistämisvaihetta syytä suorittaa geoparsing, jossa pyritään erottelemaan geokoodattavat paikannimet tekstistä. Geoparsing voidaan suorittaa käyttämällä sanastoa, jolla suodatetaan pois tavalliset sanat, käyttämällä sääntöjä paikannimien erottamiselle tai hyödyntämällä koneopetusta.

Yhdistämisalgoritmit voidaan toteuttaa monella eri tapaa. Tässä työssä tarkastellut menetelmät ovat Levenshteinin etäisyys, Longest common subsequence ja n-grammi. Näistä trigrammi, eli n-grammin versio, todettiin kokonaistuloksen osalta parhaaksi vaihtoehdoksi. Trigrammi tuotti parhaan tarkkuuden tekstipaikannuksen vertailussa ja sen myötä myös yhteiskäytön vertailussa. Lisäksi se tuotti vähiten virheellisiä osumia ja oli testatuista menetelmistä suoritusajaltaan nopein. Puutteiden määrä trigrammilla geokoodattaessa oli hieman suurempi kuin muilla menetelmillä, mutta ero ei ole huomattava. Menetelmän valintaan kuitenkin vaikuttaa myös geokoodauksen tarkoitus. Jos on mahdollista käydä läpi saadut tulokset manuaalisesti ja halutaan saada mahdollisimman paljon oikeita osumia, saattaa olla parempi käyttää LCS vertailua hyödyntävää geokoodausta, vaikka se tuottaakin eniten virheellisiä tuloksia. Työn tuloksista voidaan havaita, että geokoodaustyökalujen antamat tulokset ja niiden tarkkuus kuitenkin vaihtelevat erilaisten aineistojen välillä. Tästä syystä tämän tutkimuksen pohjalta ei voida vetää kovin pitkälle meneviä johtopäätöksiä siitä kuinka tarkkoja tässä esitetyt menetelmät ovat. Tavallisten sanojen paikantamisen väheneminen geoparsingin parantumisen myötä saattaa johtaa siihen, että tarkkuussuhteet eri yhdistämisalgoritmeja käyttävien menetelmien välillä muuttuvat.

Geokoodaus on jatkuvaa tasapainoilua oikeiden ja virheellisten tulosten määrien sekä suoritusajan välillä. Menetelmä joka tuottaa enemmän oikeita tuloksia tuottaa väistämättä myös enemmän vääriä tuloksia. Jos geokoodauksessa tehtäville vertailuille asetetaan löysät kriteerit, eli suoritetaan vertailu kaikkiin sanoihin niiden pituudesta ja alkukirjaimesta huolimatta, saadaan enemmän sekä oikeita että virheellisisä osumia kuin tiukoilla kriteereillä ja samalla suoritusaika pitenee. Tässä työssä käytetty haun rajaaminen kotimaan perusteella, johtaa siihen, ettei kaikkia selkeitä paikkailmaisuja onnistuta paikantamaan, mutta toisaalta se vähentää virheellisten paikannusten määrää huomattavasti ja samalla nopeutta prosessin suoritusta.

Jotta geokoodauksen tarkkuutta saadaan parannettua tässä työssä kehitettyjen työkalujen tarkkuudesta, on ensisijaisesti pystyttävä rajoittamaan tavallisten sanojen paikantamista. Tekstistä on pystyttävä erottamaan sanat jotka saattavat olla paikannimiä käyttämällä tehokkaampaa geoparsingmenetelmää kuin tässä käytetty hukkasanalista. Mahdollisia vaihtoehtoja ovat sääntöpohjainen geoparsing ja koneoppimisen hyödyntäminen. Lisäksi vertailuaineiston tietorakennetta ja laajuutta on syytä parantaa nopeamman ja tarkemman tuloksen saavuttamiseksi.

Lähdeluettelo

Alex, B, Llewellyn, C, Grover, C, Oberlander, J & Tobin, R. 2016. Homing in on Twitter users: Evaluating an Enhanced Geoparser for User Profile Locations. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA). S. 3936-3944. ISBN 978-2-9517408-9-1.

Amelunxen, C. 2010. An Approach to Gecoding Based on Volunteered Spatial Data. Proceedings of Geoinformatik. Kiel, Saksa. Saatavissa: http://koenigstuhl.geog.uni-heidelberg.de/publications/2010/Amelunxen/amelunxen-geocodingOSM.pdf

Apache OpenNLP [Viitattu 20.2.2017] Saatavissa: https://opennlp.apache.org/index.html

Cayo, M.R. and Talbot, T.O. 2003. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics. Vol. 2:1. 10 s. Saatavissa: http://www.ij-healthgeographics.com/content/2/1/10

Charif O. ym. 2010. A method and a tool for geocoding and record linkage. Geosience and Remote Sensing (IITA-GRS). 8 s. IBSN 978-1-4244-8517-8 (sähköinen)

Christen ym. 2004. A Probabilistic Geocoding System based on a National Address File. Proceedings of the 3rd Australasian Data Mining Conference. 13 s. Saatavissa: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.543&rep=rep1&type=pdf

Clavin. 2015. [Viitattu 20.2.2017] Saatavissa: https://clavin.bericotechnologies.com/clavin-core/

Dredze M., Osborne M. ja Kambadur P. 2016. Geolocation for Twitter: Timing Matters. Proseedings of NAACL-HLT. Sandiego, California. S. 1064-1069. Saatavissaa: https://www.cs.jhu.edu/~mdredze/publications/2016_naacl_tweet_geolocation.pdf

Freire N. ym. 2011. A Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records. Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. Ottawa, Ontario, Canada. S. 339-348. Saatavissa: http://algos.inesc-id.pt/~jpa/InscI/poisson/varwwwhtml/portal/ficheiros/publicacoes/9653.pdf

GeoNames http://www.geonames.org/ [30.1.2017]

Goldberg ym., 2007 From Text to Geographic Coordinates: The Current State of Geocoding. URISA Journal Vol.19:1 S.33-46. Saatavissa: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.3589&rep=rep1&type=pdf#page=34

Goldberg, D 2008 A Geocoding Best Practices Guide. Springfield, IL. North American Association of Central Cancer Registries. 287 s. Saatavissa: https://www.naaccr.org/LinkClick.aspx?fileticket=ZKekM8k_IQ0%3D&tabid=239&mid=699

Goldberg, D. & Cockburn M. 2010. Improving Geocode Accuracy with Candidate Selection Criteria. Transactions in GIS. Vol. 14:1. S.149-176.

Goldberg, D. 2011. Advances in Geocoding Research and Practice. Transactions in GIS. Vol. 15:6. S.727-733.

Goldberg D. 2013. Geocoding Techniques and Technologies for Location-Based Services. Teoksessa: H. Karimi. Advance location-based technologies and services. S.75-106. Saatavissa: https://ebooks-it.org/1466518189-ebook.htm IBSN 978-1-4665-1818-6 (painettu) IBSN 978-1-4665-1819-3 (sähköinen)

Han B., Cook P. & Baldwin T. 2014. Text-Based Twitter User Geolocation Prediction. Journal of Artificial Intelligence Research. Vol. 49:1. S. 451-500. Saatavissa: http://www.jair.org/media/4200/live-4200-7781-jair.pdf

Järvelin A., Järvelin A. & Järvelin K. 2007. s-grams: Defining generalized n-grams for information retrieval. Information Processing & Management. Vol. 43:4. S. 1005-1019. Saatavissa: https://ccc.inaoep.mx/~villasen/bib/s-grams%20-%20defininf%20generalized%20n-grams%20for%20information%20retrieval.pdf

Leidner J. & Lieberman M. 2011. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. SIGSPATIAL Special. Vol. 3:2. S. 5-11. Saatavissa: http://www.sigspatial.org/sigspatial-special-issues/SIGSPATIALSpecial-Vol3Num2Jul2011.pdf#page=9

Lennert M. 2015. The Use of Exhaustive Micro-Data Firm Databases for Economic Geography: The Issues of Geocoding and Usability in the Case of the Amadeus Database. ISPRS International Journal Geo-Information. Vol. 4:1. S. 62-86. ISSN 2220-9964

LingPipe [Viitattu 20.2.2017] Saatavissa: http://ir.exp.sis.pitt.edu/ne/lingpipe-2.4.0/index.html

Maanmittauslaitos. Maastotietokannan osoitteiden kyselypalvelu (WFS). [Viitattu 28.3.2017] Saatavissa: http://www.maanmittauslaitos.fi/kartat-ja-paikkatieto/asiantuntevalle-kayttajalle/kartta-ja-paikkatietojen-rajapintapalvelut-4

Nikita. 2011. Fuzzy string search. Nikitas's blog. [Viitattu 1.2.2017]. Saatavissa: http://ntz-develop.blogspot.fi/2011/03/fuzzy-string-search.html

Ranks NL. Stopwords. [Viitattu 20.2.2017] Saatavissa: http://www.ranks.nl/stopwords

Ranzijn B. 2013. A Geocoding Algorithm Based On A Comparative Study Of Address Matching Techniques. Masters Thesis. Erasmus Universiteit, Operations Research and Quantitative Logistics. Rotterdam.

Recchia G. ja Louwerse M. 2013. A Comparison of String Similarity Measures for Toponym Matching. COMP@ SIGSPATIAL. Orlando FL, USA. S. 54-61. Saatavissa: https://pdfs.semanticscholar.org/15db/1f773eaec954868187330affa1f58a270552.pdf ISBN 978-1-4503-2535-6

Roongpiboonsopit D. & Karimi H. 2010. Comparative evaluation and analysis of online geocoding services. International Journal of Geographical Information Science. Vol. 24:7. S. 1081-1100.

Sidkar U. ja Gambäck B. 2016. Feature-Rich Twitter Named Entity Recognition and Classification. Proceedings of the 2nd Workshop on Noisy User-generated Text. S. 164-170. Osaka, Japan. Saatavissa: https://noisy-text.github.io/2016/pdf/WNUT22.pdf

Thomson Reuters. OpenCalais [Viitattu 20.2.2017] Saatavissa: http://www.opencalais.com/
Tieteen termipankki (2017) [Viitattu 20.2.2017] Saatavissa http://tieteentermipankki.fi/wiki/Termipankki:Etusivu

Zandbergen P. 2008. A comparison of address point, parcel and street geocoding techniques. Computers, Environment and Urban Systems. Vol. 32:3. S. 214-232.

Zandbergen, P. 2009. Geocoding Quality and Implications for Spatial Analysis. Geography Compass. Vol. 3:2. S. 647-680. Saatavissa: www.academia.edu/download/31455419/Zandbergen_Geography_Compass_2009.pdf

Zhang W. & Gelernter J. 2013. Cross-lingual geo-parsing for non-structured data. Proc. 7th Workshop on Geographic Information Retrieval. New York. S. 64–71.

Zhang W. & Gelernter J. 2014. Geocoding location expressions in Twitter messages: A preference learning method. Journal of Spatial Information Science. Vol 9. S. 37-70. Saatavissa: http://josis.org/index.php/josis/article/view/170/129

Liiteluettelo

Liite 1. Tulokset

Hiihtoloma-aineiston geokoodauksen tulokset.

Yhteiset	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Paikka	105	95.5	105	100.0	0	0.0	0	5
Kotipaikka	449	66.4	426	94.9	13	2.90	23	49
Kotimaa	587	86.8	578	98.47			9	36

Levenshtein (0.25/0.17)	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Teksti	242	26.6	137	56.6	41	16.9	105	134
tarkka	181	19.9	128	70.72	31	17.1	53	
epätarkka	61	6.7	9	14.8	10	16.4	52	
Teksti 2	57	25.3	9	15.8	26	45.6	48	
Yhdessä (ei teksti 2)	593	65.1	489	82.5	43	7.25	104	65
Yhdessä (kaikki)	647	71.0	498	77.0	69	10.66	149	51

Levenshtein (0.17/0.17)	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Teksti	207	22.7	135	65.2	36	17.4	72	136
tarkka	184	20.2	130	70.65	32	17.4	54	
epätarkka	23	2.5	5	21.7	4	17.4	18	
Teksti 2	57	25.3	9	15.8	26	45.6	48	
Yhdessä (ei teksti 2)	584	64.1	508	87.0	40	6.85	76	67
Yhdessä (kaikki)	638	70.0	517	81.0	66	10.34	121	52

LCS	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Teksti	318	34.9	147	46.2	96	30.2	171	124
tarkka	166	18.2	113	68.1	31	18.7	53	
epätarkka	152	16.7	34	22.4	65	42.8	118	
Teksti 2	97	43.1	9	9.3	73	75.3	88	
Yhdessä (ei teksti 2)	612	67.2	449	73.4	85	13.89	163	59
Yhdessä (kaikki)	702	77.1	458	65.2	153	21.79	244	35

Digram (N=2)	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Teksti	217	23.8	141	65.0	43	19.8	76	130
tarkka	185	20.3	130	70.3	35	18.9	55	
epätarkka	32	3.5	11	34.4	8	25.0	21	
Teksti 2	69	30.7	10	14.5	43	62.3	59	
Yhdessä (ei teksti 2)	585	64.2	508	86.8	43	7.35	77	66
Yhdessä (kaikki)	647	71.0	518	80.1	82	12.67	129	47

Trigram (N=3)	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Teksti	189	20.7	135	71.4	33	17.5	54	136
tarkka	183	20.1	130	71.0	32	17.5	53	
epätarkka	6	0.7	5	83.3	1	16.7	1	
Teksti 2	19	8.4	4	21.1	9	47.4	15	
Yhdessä (ei teksti 2)	581	63.8	520	89.5	36	6.20	61	69
Yhdessä (kaikki)	605	66.4	525	86.8	45	7.44	80	59

Pakkoruotsi-aineiston geokoodauksen tulokset.

Yhteiset	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Paikka	29	100.0	29	100.0	0	0.0	0	0
Kotipaikka	265	37.5	264	99.6	16	6.04	1	8
Kotimaa	646	91.4	645	99.85			1	15

Levenshtein (0.17)	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Teksti	130	12.4	2	1.5	94	72.3	128	8
tarkka	105	10.0	1	1.0	78	74.3	104	
epätarkka	25	2.4	1	4.0	16	64.0	24	
Teksti 2	85	28.1	0	0.0	32	37.6	85	
Yhdessä (ei teksti 2)	394	37.5	266	67.5	109	27.66	128	6
Yhdessä (kaikki)	479	45.6	266	55.5	141	29.44	213	5

Digram (N=2)	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Teksti	165	15.7	3	1.8	94	57.0	162	7
tarkka	105	10.0	1	1.0	78	74.3	104	
epätarkka	60	5.7	2	3.3	16	26.7	58	
Teksti 2	94	31.1	0	0.0	70	74.5	94	
Yhdessä (ei teksti 2)	428	40.7	266	62.1	109	25.47	162	6
Yhdessä (kaikki)	522	49.7	266	51.0	179	34.29	256	4

Trigram (N=3)	Osuma-osuus, kpl	Osuma-osuus, %	Oikein, kpl	Tarkkuus, %	Useita, kpl	Useita, %	Virhe, kpl	Puute, kpl
Teksti	106	10.1	1	0.9	78	73.6	105	9
tarkka	105	10.0	1	1.0	78	74.3	104	
epätarkka	1	0.1	0	0.0	0	0.0	1	
Teksti 2	0	0.0	0	0.0	0	0.0	0	
Yhdessä (ei teksti 2)	374	35.6	268	71.7	93	24.87	106	6
Yhdessä (kaikki)	401	38.2	268	66.8	110	27.43	133	5

Liite 2. Suoritusajat

Hiihtoloma-aineiston geokoodauksen suoritusajat.

Suoritusajat	Levenshtein (0.25/0.17)	Levenshtein (0.17/0.17)	LCS	Digram (N=2)	Trigram (N=3)
Paikka	48,89 sek				
Teksti	27 min 32 sek	28 min 10 sek	27 min 15 sek	27 min 21 sek	27 min 37 sek
Teksti 2	23 min 28 sek	23 min 43 sek	17 min 24 sek	13 min 14 sek	13 min 29 sek
Kotipaikka	2 min 19 sek				
Yhdessä (ei teksti 2)	27 min	26 min 27 sek	26 min 44 sek	26 min 21 sek	26 min 8 sek
Yhdessä (kaikki)	48 min 44 sek	48 min 55 sek	43 min 28 sek	38 min 50 sek	39 min 7 sek

Pakkoruotsi-aineiston geokoodauksen suoritusajat.

Suoritusajat	Levenshtein (0.17/0.17)	Digram (N=2)	Trigram (N=3)
Paikka	42,91 sek		
Teksti	39 min 20 sek	37 min 55 sek	38 min 49 sek
Teksti 2	38 min 58 sek	20 min 15 sek	33 min 57
Kotipaikka	1 min 17 sek		
Yhdessä (ei teksti 2)	39 min 37 sek	39 min 16 sek	38 min 49 sek
Yhdessä (kaikki)	1 h 18 min 5 sek	59 min 14 sek	1 h 51 sek

Liite 3. Tulokset tasoittain

Hiihtoloma-aineiston yhteisgeokoodauksen tulos tasoittain

Levenshtein (0.25/0.17)				
Taso	Osuus, kpl	Osuus, %	Oikein, kpl	Tarkkuus, %
1	24	3.71	24	100
2	82	12.67	82	100
25	0	0	0	0
3	116	17.93	87	75.0
35	26	4.02	5	19.2
4	42	6.49	5	11.9
45	9	1.39	2	22.2
5	286	44.20	276	96.5
55	8	1.24	8	100.0
6	28	4.33	5	17.9
65	26	4.02	4	15.4
yht	647	100.00	498	77.0

Levenshtein (0.17/0.17)				
Taso	Osuus, kpl	Osuus, %	Oikein, kpl	Tarkkuus, %
1	24	3.76	24	100
2	82	12.85	82	100
25	0	0	0	0
3	117	18.34	87	74.4
35	27	4.23	7	25.9
4	17	2.66	5	29.4
45	4	0.63	0	0
5	304	47.65	294	96.7
55	9	1.41	9	100
6	28	4.39	5	17.86
65	26	4.08	4	15.38
yht	638	100.00	517	81.0

LCS				
Taso	Osuus, kpl	Osuus, %	Oikein, kpl	Tarkkuus, %
1	24	3.42	24	100
2	82	11.68	82	100
25	0	0	0	0
3	108	15.38	80	74.1
35	26	3.70	6	23.1
4	77	10.97	11	14.3
45	52	7.41	13	25.0
5	236	33.62	226	95.8
55	7	1.00	7	100.0
6	22	3.13	2	9.1
65	68	9.69	7	10.3
yht	702	100.00	458	65.2

Digram (N=2)				
Taso	Osuus, kpl	Osuus, %	Oikein, kpl	Tarkkuus, %
1	24	3.71	24	100
2	82	12.67	82	100
25	0	0	0	0
3	116	17.93	88	75.9
35	27	4.17	7	25.9
4	22	3.40	9	40.9
45	7	1.08	1	14.3
5	298	46.06	288	96.6
55	9	1.39	9	100.0
6	23	3.55	4	17.4
65	39	6.03	6	15.4
yht	647	100.00	518	80.1

Trigram (N=3)				
Taso	Osuus, kpl	Osuus, %	Oikein, kpl	Tarkkuus, %
1	24	3.97	24	100
2	82	13.55	82	100
25	0	0	0	0
3	117	19.34	88	75.2
35	27	4.46	7	25.9
4	5	0.83	4	80.0
45	0	0	0	0
5	317	52.40	307	96.8
55	9	1.49	9	100.0
6	15	2.48	4	26.7
65	9	1.49	0	0
yht	605	100.00	525	86.8

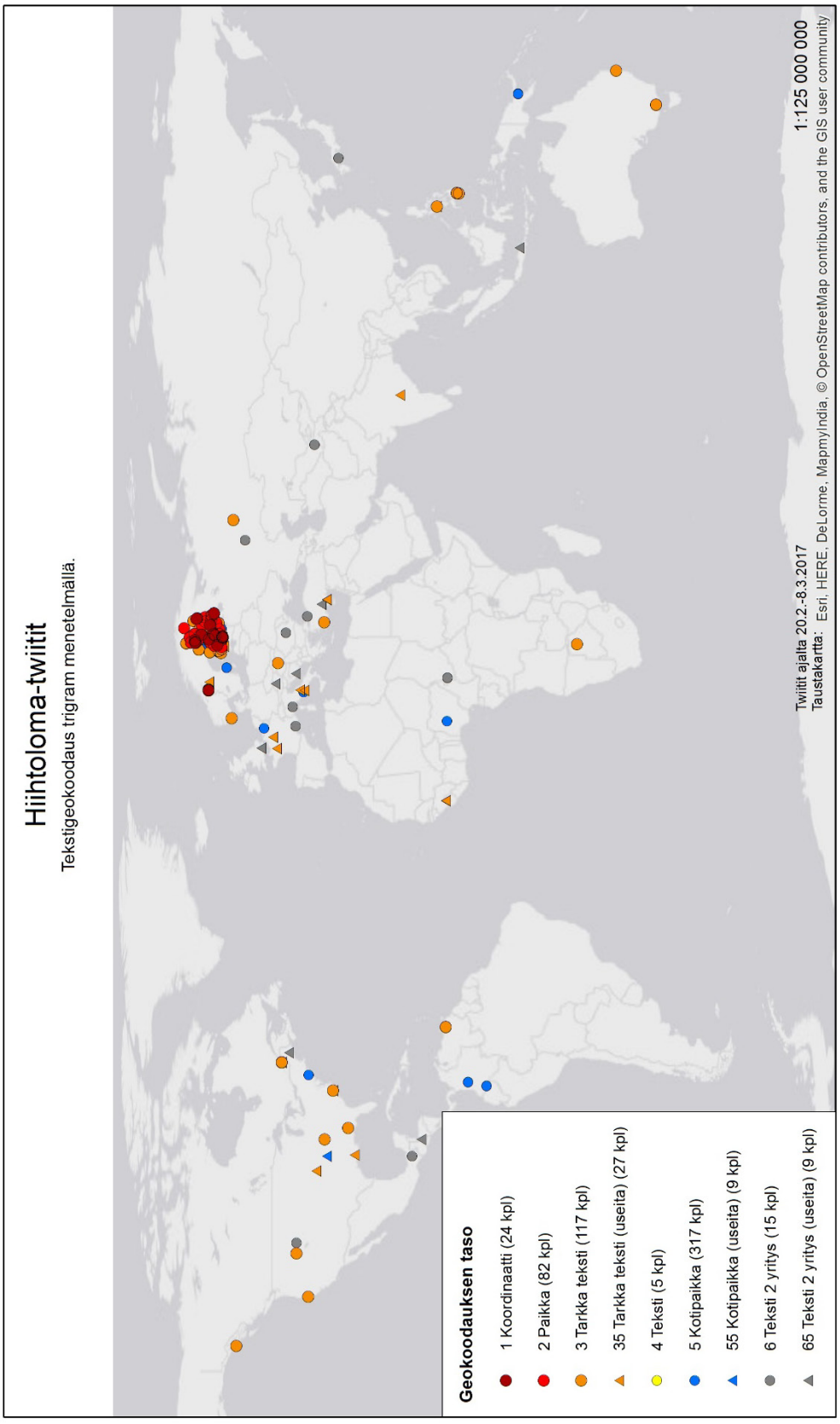
Pakkoruotsi-aineiston geokoodauksen tulos tasoittain.

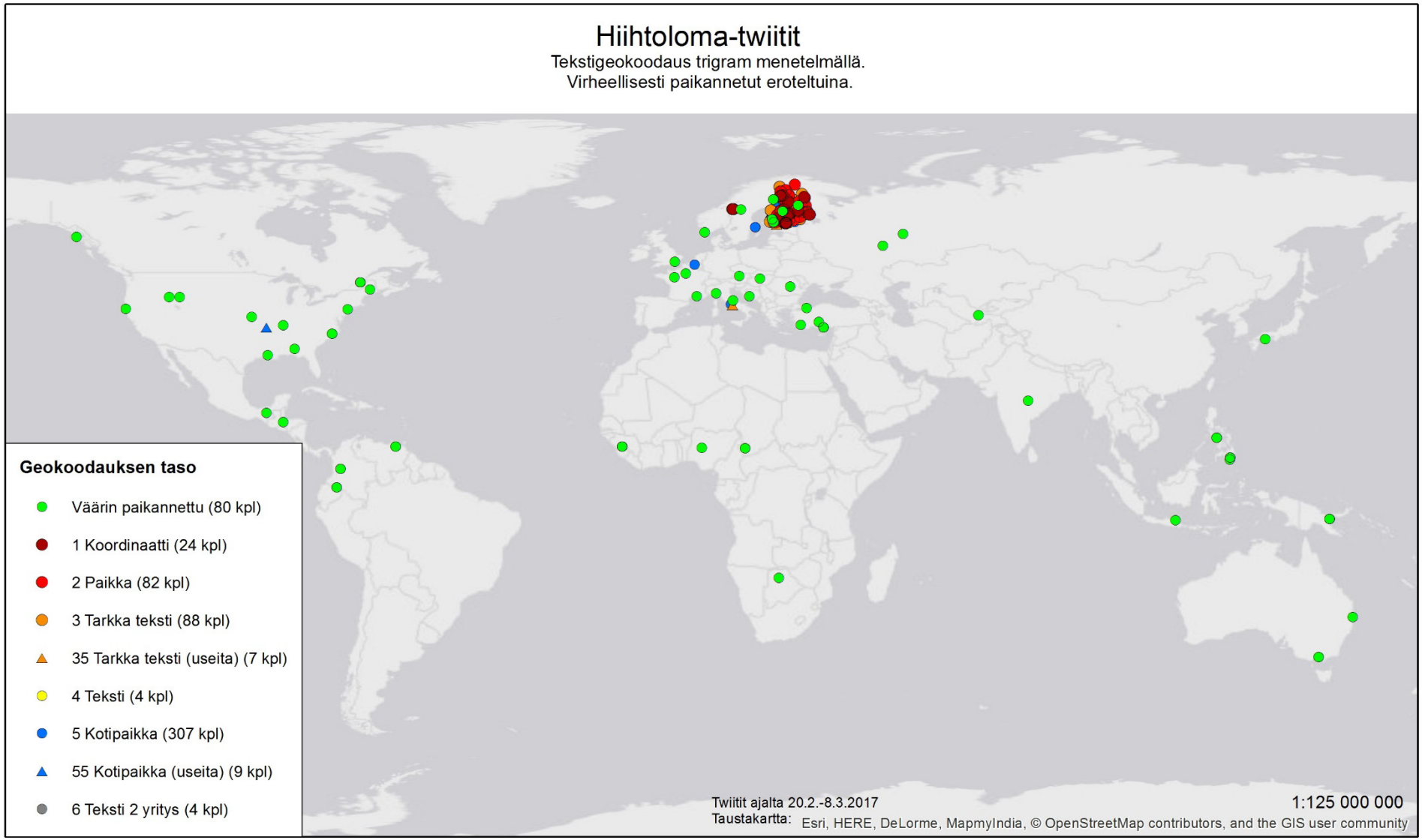
Levenshtein (0.17/0.17)				
Taso	Osuus, kpl	Osuus, %	Oikein, kpl	Tarkkuus, %
1	0	0.00	0	0
2	29	6.05	29	100
25	0	0.00	0	0
3	27	5.64	1	3.7
35	78	16.28	0	0.0
4	8	1.67	1	12.5
45	16	3.34	0	0
5	221	46.14	220	99.5
55	15	3.13	15	100.0
6	53	11.06	0	0
65	32	6.68	0	0
yht	479	100.00	266	55.5

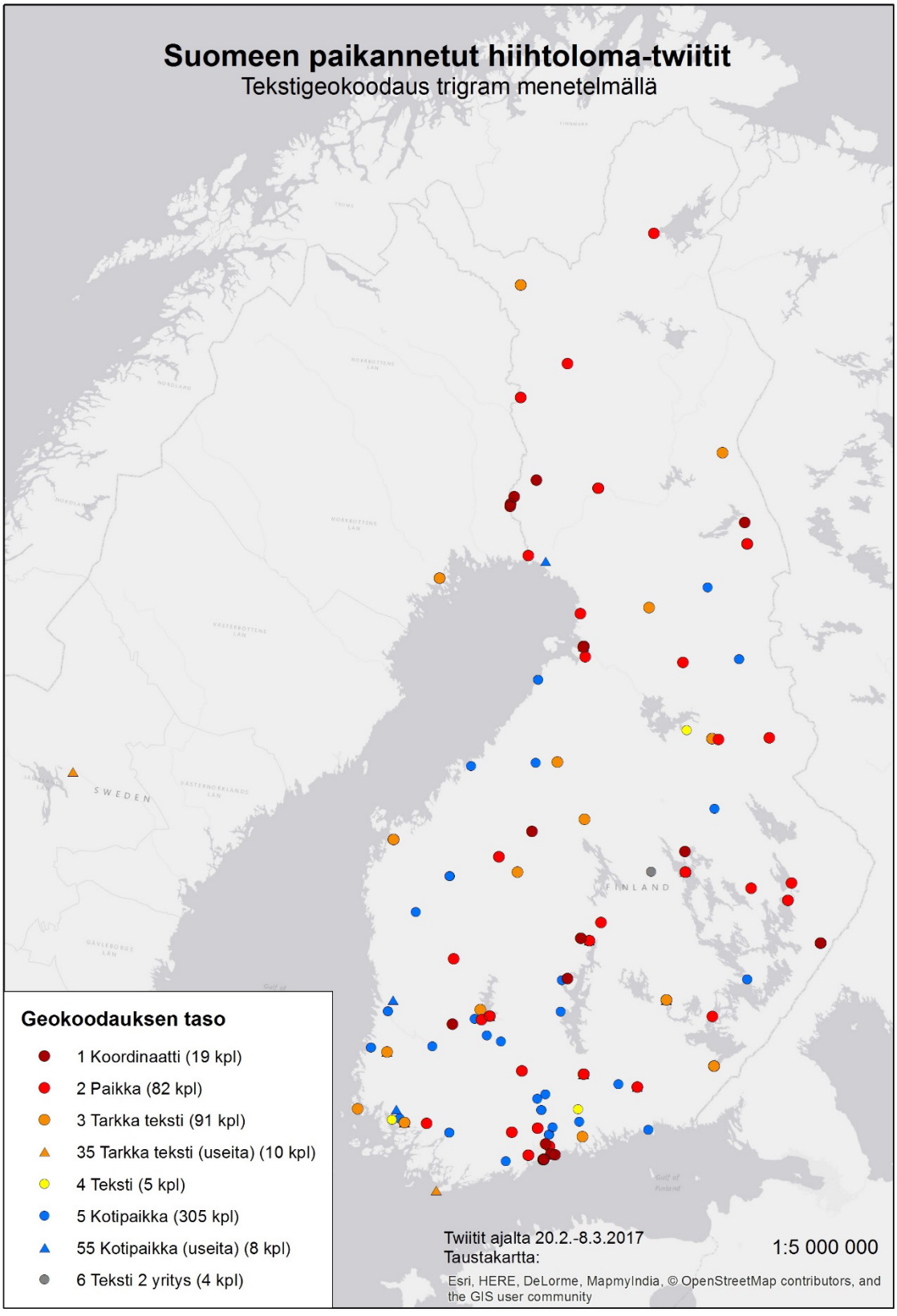
Digram (N=2)				
Taso	Osuus, kpl	Osuus, %	Oikein, kpl	Tarkkuus, %
1	0	0.00	0	0
2	29	5.56	29	100
25	0	0.00	0	0
3	27	5.17	1	3.7
35	78	14.94	0	0.0
4	43	8.24	2	4.7
45	16	3.07	0	0
5	220	42.15	219	99.5
55	15	2.87	15	100.0
6	24	4.60	0	0
65	70	13.41	0	0
yht	522	100.00	266	51.0

Trigram (N=3)				
Taso	Osuus, kpl	Osuus, %	Oikein, kpl	Tarkkuus, %
1	0	0.00	0	0
2	29	7.23	29	100.0
25	0	0.00	0	0
3	27	6.73	1	3.7
35	78	19.45	0	0
4	1	0.25	0	0
45	0	0.00	0	0
5	224	55.86	223	99.6
55	15	3.74	15	100.0
6	10	2.49	0	0
65	17	4.24	0	0
yht	401	100.00	268	66.8

Liite 4. Kartat




Hiihtoloma-twiitit
Tekstigeokoodaus trigram menetelmällä.
Virheellisesti paikannetut eroteltuina.
Geokoodauksen taso
Väärin paikannettu (80 kpl)
1 Koordinaatti (24 kpl)
2 Paikka (82 kpl)
3 Tarkka teksti (88 kpl)
35 Tarkka teksti (useita) (7 kpl)
4 Teksti (4 kpl)
5 Kotipaikka (307 kpl)
55 Kotipaikka (useita) (9 kpl)
6 Teksti 2 yritys (4 kpl)
Twiitit ajalta 20.2.-8.3.2017
1:125 000 000
Taustakartta: Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community


Suomeen paikannetut hiihtoloma-twiitit
Tekstigeokoodaus trigram menetelmällä
SWEDEN
FINLAND
Geokoodauksen taso
1 Koordinaatti (19 kpl)
2 Paikka (82 kpl)
3 Tarkka teksti (91 kpl)
35 Tarkka teksti (useita) (10 kpl)
4 Teksti (5 kpl)
5 Kotipaikka (305 kpl)
55 Kotipaikka (useita) (8 kpl)
6 Teksti 2 yritys (4 kpl)
Twiitit ajalta 20.2.-8.3.2017
Taustakartta:
1:5 000 000
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community