Master's Thesis in Geography

Geoinformatics

Developing a Finnish geoparser for extracting location information from unstructured texts

Tatu Leppämäki

2022

Supervisors: Tuuli Toivonen and Tuomo Hiippala

Master's Programme in Geography

Faculty of Science

Faculty	Department
Faculty of Science	Department of Geosciences and Geography

Author

Tatu Leppämäki

Title of thesis

Developing a Finnish geoparser for extracting location information from unstructured texts

Programme and study track

Master's programme in geography, Geoinformatics

Level of the thesis	Date	Number of pages
Master's thesis, 30 credits	March 2022	65 + 1 appendix

Abstract

Ever more data is available and shared through the internet. The big data masses often have a spatial dimension and can take many forms, one of which are digital texts, such as articles or social media posts. The geospatial links in these texts are made through place names, also called toponyms, but traditional GIS methods are unable to deal with the fuzzy linguistic information. This creates the need to transform the linguistic location information to an explicit coordinate form. Several *geoparsers* have been developed to recognize and locate toponyms in free-form texts: the task of these systems is to be a reliable source of location information. Geoparsers have been applied to topics ranging from disaster management to literary studies. Major language of study in geoparser research has been English and geoparsers tend to be language-specific, which threatens to leave the experiences provided by studying and expressed in smaller languages unexplored.

This thesis seeks to answer three research questions related to geoparsing: What are the most advanced geoparsing methods? What linguistic and geographical features complicate this multi-faceted problem? And how to evaluate the reliability and usability of geoparsers? The major contributions of this work are an open-source geoparser for Finnish texts, Finger, and two test datasets, or corpora, for testing Finnish geoparsers. One of the datasets consists of tweets and the other of news articles. All of these resources, including the relevant code for acquiring the test data and evaluating the geoparser, are shared openly.

Geoparsing can be divided into two sub-tasks: recognizing toponyms amid text flows and resolving them to the correct coordinate location. Both tasks have seen a recent turn to deep learning methods and models, where the input texts are encoded as, for example, word embeddings. Geoparsers are evaluated against gold standard datasets where toponyms and their coordinates are marked. Performance is measured on equivalence and distance-based metrics for toponym recognition and resolution respectively.

Finger uses a toponym recognition classifier built on a Finnish BERT model and a simple gazetteer query to resolve the toponyms to coordinate points. The program outputs structured geodata, with input texts and the recognized toponyms and coordinate locations. While the datasets represent different text types in terms of formality and topics, there is little difference in performance when evaluating Finger against them. The overall performance is comparable to the performance of geoparsers of English texts. Error analysis reveals multiple error sources, caused either by the inherent ambiguousness of the studied language and the geographical world or are caused by the processing itself, for example by the lemmatizer.

Finger can be improved in multiple ways, such as refining how it analyzes texts and creating more comprehensive evaluation datasets. Similarly, the geoparsing task should move towards more complex linguistic and geographical descriptions than just toponyms and coordinate points. Finger is not, in its current state, a ready source of geodata. However, the system has potential to be the first step for geoparsers for Finnish and it can be a steppingstone for future applied research.

Keywords
geoinformatics, geoparsing, toponym recognition, natural language processing, GIS, NLP, named entity recognition

Where deposited

University of Helsinki electronic theses library E-thesis/HELDA

Muita tietoja – Övriga uppgifter – Additional information

Tiedekunta	Osasto
Luonnontieteellinen tiedekunta	Geotieteiden ja maantieteen osasto

Tekijä

Tatu Leppämäki

Tutkielman otsikko

Suomenkielisen geojäsentimen kehittäminen: kuinka hankkia sijaintitietoa jäsentelemättömistä tekstiaineistoista

Koulutusohjelma ja opintosuunta

Maantieteen maisteriohjelma, geoinformatiikka

Tutkielman taso	Aika	Sivumäärä
Maisterintutkielma, 30 opintopistettä	Maaliskuu 2022	65 + 1 liite

Tiivistelmä – Referat – Abstract

Alati enemmän aineistoa tuotetaan ja jaetaan internetin kautta. Aineistot ovat vaihtelevia muodoiltaan, kuten verkkoartikkelien ja sosiaalisen media julkaisujen kaltaiset digitaaliset tekstit, ja niillä on usein spatiaalinen ulottuvuus. Teksteissä geospatiaalisuutta ilmaistaan paikannimien kautta, mutta tavanomaisilla paikkatietomenetelmillä ei kyetä käsittelemään tietoa epätäsmällisessä kielellisessä asussaan. Tämä on luonut tarpeen muuntaa tekstimuotoisen sijaintitiedon näkyvään muotoon, koordinaateiksi. Ongelmaa ratkaisemaan on kehitetty *geojäsentimiä*, jotka tunnistavat ja paikantavat paikannimet vapaista teksteistä, ja jotka oikein toimiessaan voisivat toimia paikkatiedon lähteenä maantieteellisessä tutkimuksessa. Geojäsentämistä onkin sovellettu katastrofihallinnasta kirjallisuudentutkimukseen. Merkittävässä osassa geojäsentämisen tutkimusta tutkimusaineiston kielenä on ollut englanti ja geojäsentimetkin ovat kielikohtaisia – tämä jättää pimentoon paitsi geojäsentimien kehitykseen vaikuttavat havainnot pienemmistä kielistä myös kyseisten kielten puhujien näkemykset.

Maisterintutkielmassani pyrin vastaamaan kolmeen tutkimuskysymykseen: Mitkä ovat edistyneimmät geojäsentämismenetelmät? Mitkä kielelliset ja maantieteelliset monitulkintaisuudet vaikeuttavat tämän monitahoisen ongelman ratkaisua? Ja miten arvioida geojäsentimien luotettavuutta ja käytettävyyttä? Tutkielman soveltavassa osuudessa esittelen Fingerin, geojäsentimen suomen kielelle, ja kuvaan sen kehitystä sekä suorituskyvyn arviointia. Arviointia varten loin kaksi testiaineistoa, joista toinen koostuu Twitter-julkaisuista ja toinen uutisartikkeleista. Finger-geojäsennin, testiaineistot ja relevantit ohjelmakoodit jaetaan avoimesti.

Geojäsentäminen voidaan jakaa kahteen alitehtävään: paikannimien tunnistamiseen tekstivirrasta ja paikannimien ratkaisemiseen oikeaan koordinaattipisteeseen mahdollisesti useasta kandidaatista. Molemmissa vaiheissa uusimmat metodit nojaavat syväoppimismalleihin ja -menetelmiin, joiden syötteinä ovat sanaupotusten kaltaiset vektorit. Geojäsentimien suoriutumista testataan aineistoilla, joissa paikannimet ja niiden koordinaatit tiedetään. Mittatikkuna tunnistamisessa on vastaavuus ja ratkaisemisessa etäisyys oikeasta sijainnista.

Finger käyttää paikannimitunnistinta, joka hyödyntää suomenkielistä BERT-kielimallia, ja suoraviivaista tietokantahakua paikannimien ratkaisemiseen. Ohjelmisto tuottaa taulukkomuotoiseksi jäsenneltyä paikkatietoa, joka sisältää syötetekstit ja niistä mahdollisesti tunnistetut paikannimet koordinaattisijainteineen. Testiaineistot eroavat aihepiireiltään, mutta Finger suoriutuu niillä likipitäen samoin, ja suoriutuu englanninkielisillä aineistoilla tehtyihin arviointeihin suhteutettuna kelvollisesti. Virheanalyysi paljastaa useita virhelähteitä, jotka johtuvat kielten tai maantieteellisen todellisuuden luontaisesta epäselvyydestä tai ovat prosessoinnin aiheuttamia, kuten perusmuotoistamisvirheet.

Kaikkia osia Fingerissä voidaan parantaa, muun muassa kehittämällä kielellistä käsittelyä pidemmälle ja luomalla kattavampia testiaineistoja. Samoin tulevaisuuden geojäsentimien tulee kyetä käsittelemään monimutkaisempia kielellisiä ja maantieteellisiä kuvaustapoja kuin pelkät paikannimet ja koordinaattipisteet. Finger ei nykymuodossaan tuota valmista paikkatietoa, jota kannattaisi kritiikittä käyttää. Se on kuitenkin lupaava ensiaskel suomen kielen geojäsentimille ja astinlauta vastaisuuden soveltavalle tutkimukselle.

Avainsanat
geoinformatiikka, geojäsentäminen, paikannimitunnistus, luonnollisen kielen käsittely, GIS, NLP, nimettyjen entiteettien tunnistus

Säilytyspaikka

Helsingin yliopiston sähköinen tietokanta: E-thesis/HELDA

Muita tietoja – Övriga uppgifter – Additional information

Contents

1. Introduction 1
2. Background 3
2.1. Defining geoparsing 3
2.2. Applications of geoparsing: why geoparse texts? 4
2.3. Toponyms and locations 6
2.3.1. Toponyms and linguistic ambiguity 6
2.3.2. Locations in time and space 10
2.4. Toponym recognition: Named Entity Recognition for geoparsing 13
2.4.1. Named Entity Recognition 13
2.4.2. Traditional approaches to toponym recognition 14
2.4.3. Neural NER for toponym recognition 15
2.5. Toponym resolution 17
2.5.1. Gazetteers 17
2.5.2. Toponym resolution methods 19
2.6. Geoparsing evaluation 21
2.6.1. Toponym recognition evaluation 22
2.6.2. Toponym resolution evaluation 24
2.6.3. Factors affecting geoparser performance 26
3. Data and Methods 27
3.1. Creating Finger 30
3.1.1. NLP Pipeline 31
3.1.2. GIS pipeline 33
3.1.3. Finger output 35
3.2. Creating evaluation data 36
3.3. Additional evaluation metrics: lemmatization and query errors 40
4. Results 41
4.1. Recognition and resolution evaluation on the Finger corpora 41
4.2. Error analysis 43
5. Discussion 45
5.1. Contextualizing the results 45
5.2. Developing Finger: more, and better 48
5.3. Beyond it all: next steps for geoparsing 51
6. Conclusion 53
Acknowledgements 54
Literature 55
Appendix A. Annotation practices for Finger-news and Finger-tweets 65

Table of Figures

Figure 1. Top-level view of geoparsing and an example. 3
Figure 2. A taxonomy of toponyms. 9
Figure 3. Visualizing containment and granularity differences in locations. 12
Figure 4. Example of a sentence with the named entities tagged 14
Figure 5. Example of toponym recognition evaluation. 23
Figure 6. AUC visualized 25
Figure 7. Example of toponym resolution errors and error metrics. 26
Figure 8. The workflow of the thesis. 28
Figure 9. How Finger operates. 31
Figure 10. Coverage comparison of GeoNames and NLS Placenames 35
Figure 11. Spatial characteristics of the Finger corpora 38
Figure 12. AUC error curves 43

Table of Tables

Table 1. Examples of publicly available gazetteers 18
Table 2. Input datasets and their purpose in this study 29
Table 3. Performance of the NER classifier on the internal test set.. 33
Table 4. The Finger corpora in numbers. 37
Table 5. Toponym recognition evaluation 42
Table 6. Toponym resolution evaluation.. 43
Table 7. Contextualizing Finger's performance with previous work. 47

1. Introduction

Studying human activity and physical phenomena in space through the lens of geography is being expanded and changed by new data sources: we live in the age of big data. While the volume of data is nothing new for geographical research and especially geoinformatics, which has a long history of dealing with e.g., massive satellite imagery datasets, big data is ever more varied. *Variety* refers to the messy, often unstructured nature of this data – for example, description of a holiday resort expressed in rambling travel blogs instead of gathered in structured visitor surveys (Kitchin, 2013; Miller & Goodchild, 2015; Yan et al., 2021).

These unstructured masses of data are created every day as people interact with their surroundings, which presents possibilities for geographic knowledge discovery (Miller & Goodchild, 2015). Indeed, much of all information is georeferenced directly or indirectly (Hahmann & Burghardt, 2013), waiting for geographers to exploit it. In this thesis, I focus on big collections of texts created and shared on the internet, such as social media content, news articles and Wikipedia entries, collectively named geo-text data by Hu (2018b).

Texts are linked to real-world locations by various means, such as coordinate geotags attached to social media posts and *toponyms*, also known as place names, in the texts themselves (Hu, 2018b). The latter method is more implicit because the geographic reference is expressed through language, and not in the unambiguous coordinates GI-systems operate in. This creates the need to transform the often ambiguous linguistic geoinformation in texts to an explicit form (Frank & Mark, 1991; Hu & Adams, 2021). The process of identifying toponyms and producing geographical footprints for them is called *geoparsing*. Geoparsing is further divided to two large subtasks, *toponym recognition* for identifying location mentions and *toponym resolution* for producing the correct real-world representation (Gritta, 2019; Hu & Adams, 2021).

Associating toponyms to locations is not an easy task, and has garnered research attention from fields such as geographic information retrieval (Purves et al., 2018), language technology (Gritta, 2019), and library and information science (Hill, 2006). To facilitate geoparsing, several *geoparsers* have been developed (see e.g. DeLozier et al., 2015; Karimzadeh et al., 2019; Tobin et al., 2010) and they perform well at least on some text genres (J. Wang & Hu, 2019a). Many geoparsers are language-specific and cover a limited range of languages, most prominently English and other Indo-European languages.

As Bender (2019) argues, English does not epitomize all natural languages. Even in online spaces, where English is dominant, there is significant diversity and prominence of local languages to be found (Hiippala et al., 2020). Ignoring tool development for smaller languages would mean to ignore the insights and experiences the speakers of those languages provide. In addition, the unique problems and possibilities of other languages, such as the rich morphological inflection present in Finnish but not in English, may get overlooked when the pool of researched languages is small. While recent Master's theses worked on associating Finnish war time documents (Heino, 2017) and Finnish tweets related to sports facilities (Koivisto, 2021) to locations, these solutions were purpose-built. To my knowledge, no general-purpose geoparser for Finnish texts exist.

Against this background, I first aim to explore how this multifaceted problem has been approached from the point of view of geography and language technology. Knowledge from the literature review is used in developing **Finger**, the first open-source, general-purpose geoparser for Finnish. Lastly, I test the validity of the system, discuss its strengths and weaknesses, and point out possible research directions. I seek to answer the following research questions:

RQ 1. What are the current state-of-the-art methods for recognizing and resolving toponyms?

RQ 2. What problems emerge linguistically and geographically when tackling geoparsing?

RQ 3. How reliable and usable is Finger and its output?

I answer RQ1 in Chapter 2 of the thesis by the way of a presenting recent literature on mostly English language geoparsing research. Chapter 2 progresses from the definition of geoparsing to prominent applications. These sections lay the groundwork for the exploration of toponym recognition (Section 2.4) and resolution (Section 2.5) methods. RQ 2 is addressed in Section 2.3, where the ambiguity of linguistic and coordinate-based representations of the world is explored. Finally, RQ 3 addresses *evaluation* of Finger, the geoparser developed in this work. The methodical background of geoparser evaluation is explored in Section 2.6, which is then applied to evaluating Finger in Chapter 4 and contrasting the results to previous research and discussing the geoparser's reliability in Section 5.1.

2. Background

2.1. Defining geoparsing

The central concept of this thesis, *geoparsing*, has at least two definitions that I am aware of. First, among others, Gritta et al. (2017) and Hu & Adams (2021) use geoparsing to refer to the two-step process of recognizing toponyms (toponym recognition also known as *geotagging*) in unstructured texts and resolving them unambiguously to a geographical location (toponym resolution, also known as *geocoding*). Second, for example Purves et al. (2018) and Monteiro et al. (2016) use *geoparsing* to refer only to the first part of the process, toponym recognition. In this thesis, I adopt the first definition: geoparsing is finding toponyms in texts and representing them geographically. See Figure 1 for a summarization of the major tasks in geoparsing and an example sentence run through an imaginary geoparser.

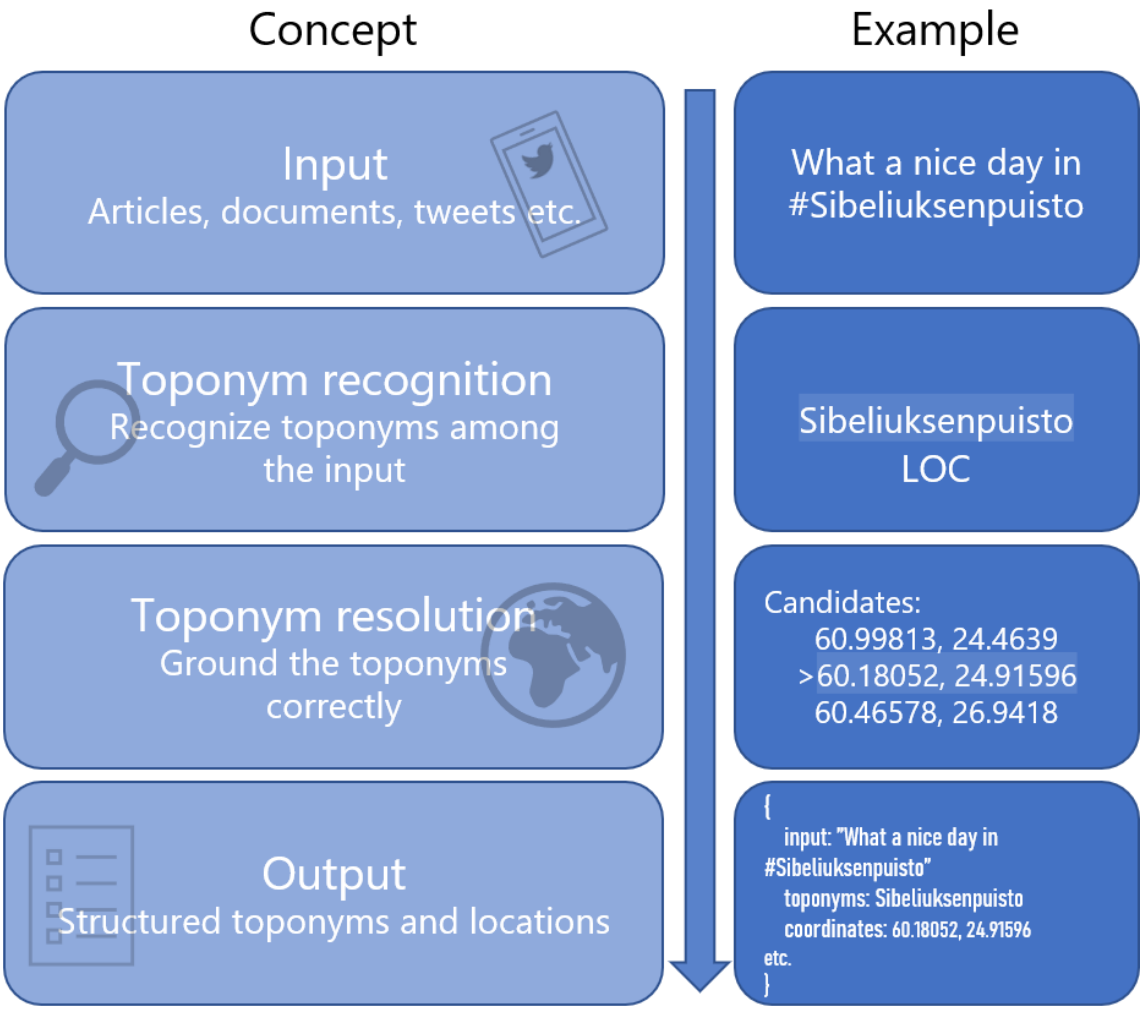


Figure 1. Top-level view of geoparsing and an example.

To clarify on some of the terminology and choices: toponym recognition is a cumbersome term, but I use it over geotagging to avoid confusion with the act of tagging online posts with locations, also called geotagging. The more general process of linking documents or individual place references to locations *georeferencing* (Purves et al., 2018, p. 206); geoparsing is one method to achieve this. *Geoparser* is a system that runs such a process on unstructured texts (Karimzadeh et al., 2019). Document, in this context, refers to any span of text forming an identifiable whole, such as a news article or a tweet. Unstructured, or free, text is distinguished from structured text by the fact that its linguistic content and form are unknown beforehand. For example, a list of addresses is structured whereas a stream of online comments is unstructured text.

In an ideal scenario, the geoparser in Figure 1 will tag *Sibeliuksenpuisto* and resolve it e.g. to a WGS84 coordinate pair (60.18052, 24.91596) within the park. This highlights a one more critical distinction in the objective of geoparsing. The goal might either be to locate each place name or to locate a geographical footprint, or the theme, of a whole document (DeLozier et al., 2016; Melo & Martins, 2017). To phrase it another way, sometimes the task is to geolocate each toponym in a document and leave further analysis to the user; sometimes to synthesize the toponym mentions and find a general target area the document discusses. These adjacent tasks can be named *document* and *toponym* geoparsing (Gritta et al., 2018). In this thesis, my emphasis is on geoparsing of toponyms. More information on document geoparsing can be found in recent reviews (Melo & Martins, 2017; Purves et al., 2018).

2.2. Applications of geoparsing: why geoparse texts?

Geoparsing has several use cases for texts on the internet (see e.g. Middleton et al., 2018). Because social media platforms, especially Twitter, provide a constant feed of content, there have been efforts to monitor events first emerging online, especially to extract information for disaster response: which areas are most severely affected, where to direct aid? (J. Wang et al., 2020). Gelernter and Mushegian (2011) did an early exploration of using Twitter for disaster relief by geoparsing tweets related to an earthquake in New Zealand. Work in this direction is expanded by Avvenuti et al. (2018), who present a geoparser as part of their *CrisMap* system. The system geoparses tweets, which are then content analyzed and mapped. Eilander et al. (2016) combined tweets located by matching neighborhood names with geodata on e.g. elevation to map flood observations. Recently, Koivisto (2021) explored

the use of sports facilities as revealed by geotags and geoparsed Twitter posts. Beyond social media, online news articles have been geoparsed to monitor disease spread (Gritta, 2019, p. 95).

Geoparsers can benefit many fields dealing with textual documents and the geospatial – what Gregory et al. (2015) call spatial humanities – by grounding the locations expressed in narratives. This allows possibilities from simple exploration to finding links between locations or any succeeding analysis. For example, Moncla et al. (2019) geoparsed and visualized Parisian urban roads mentioned in French novels to contextualize them spatially. A similar work was carried out for historical and contemporary literature set in the City of Edinburgh. The results are viewable online [1] (Alex et al., 2019). For a historical application, see the three examples presented in Alex et al. (2015): understanding the effects of 19th century commodity trade, and geoparsing ancient and historical texts. Another perspective is by Tateosian et al (2017), who tracked and mapped the spread of potato famine in the mid-19th century based on mentions in documents and literature of the time. In Finnish, Heino (2017) linked Finnish war-time documents to locations.

The field of geographic information retrieval has approached geoparsing from a document indexing and retrieval point-of-view. If a user for example queries a search engine for *hiking paths near Muonio*, a good system would be able return relevant results even if they are in neighboring regions because this info has been georeferenced and spatially indexed. A comprehensive review into this direction is given by Purves et al. (2018). As a specific example, scientific articles could benefit from being indexed not just thematically but spatially (Karl, 2019).

The above listing is not exhausting but it suffices to highlight a few points. First, there is a need for geoparsers in multiple fields of study and the need is only heightened by the explosion of internet texts and digital libraries. Second, geoparsing is not the goal, but rather a means to an end; a step in pipelines consisting of data retrieval, geoparsing, geospatial analysis and visualization (Hu, 2018b, p. 11). This does not lessen the importance of geoparsers: following from the principle of garbage in, garbage out, the geoparser output must be reliable for the subsequent steps to be reliable.

However, as argued by Hu and Adams (2021), it is important to ask: why do we need specifically geoparsed information? Certainly it is true that for example the social media datasets are plagued by problems like being unrepresentative of the general population, as they might be self-selecting instead (Miller & Goodchild, 2015). Nonetheless, Hu and Adams (2021) list several benefits of geoparsed data: because the input is freeform texts, such as travel blog entries, they open a window into people's

[1] www.litlong.org

experiences of places. These datasets also might not be spatially usable in any other way, and the data on social media is produced at a velocity that far outpaces other sources, which allows almost real-time monitoring of events (Hu & Adams, 2021, p. 2).

Finally, what can correctly located toponyms be used for; what sort of information do they tell and what not? When the information is user-provided, meaning that it has an author, it can be distinguished whether the data is *from* somewhere or *about* somewhere (Hu, 2018b; MacEachren et al., 2011; Zheng et al., 2018). The first implies that the user is or was physically present somewhere and the second that it is merely the topic of discussion, e.g. "What a sunny day here in Turku [from], I bet it's not as nice in Tampere! [about]". The distinction is relevant when, for example, studying the presence of people: unlike a geotag often does, a toponym mention does not in itself indicate a user did an action like a national park visit (Heikinheimo et al., 2017).

2.3. Toponyms and locations

Among the central concepts of this thesis are *location* and *toponym*: they are elaborated on in this chapter. In this thesis, I draw from the definitions given by Gritta et al. (2020, p. 690). A location is one of innumerable spots on Earth defined unambiguously by spatial reference systems, often represented by vector-based primitives such as points or polygons. Toponyms are named entities that label a particular location. This definition is close to the classic concepts of space and place (Tuan, 1979): thus, locations exist in clearly defined coordinate spaces, while toponyms inhibit a platial world, only existing because of humans who give them meaning (Blaschke et al., 2018). While the described definition is adopted in this thesis, see Purves and Derungs (2015) for a more nuanced discussion, where the authors aim to go beyond the "reduction of place to a name and a set of coordinates" (p. 77).

2.3.1. Toponyms and linguistic ambiguity

Ambiguity, or multiple interpretations of words and phrases, is an inseparable feature of natural languages (Pilehvar & Camacho-Collados, 2020, pp. 1–2). To be clear, natural languages are simply human languages – as separate from artificial languages, such as the programming language Python. Ambiguity can express itself in many ways, such as words that have multiple meanings depending on the context, like *bat* in English (lexical ambiguity) or how a concept is replaced with a similar one,

like using *Arkadianmäki* to represent the Finnish state powers due to the location of the parliament building (metonymic ambiguity) (Pilehvar & Camacho-Collados, 2020, pp. 1–2).

Amitay et al. (2004) defined the ambiguities plaguing geoparsing by naming geo/geo and geo/non-geo ambiguities. Geo/geo occurs when multiple locations share a name, such as Helsinki, the Capital of Finland and Helsinki, a village in Southwest Finland. Geo/non-geo is when a non-geographic entity shares the same surface-form as the location: for example, *Lahti* is a surname, a geographic formation (*bay*) and a city in southern Finland. It is the task of a geoparser to discern the ambiguities: methods to disambiguate them are explored in Sections 2.4–2.5. However, Gritta et al. (2020) go beyond by problematizing how toponyms have been used in previous research, and argue the current definition is under-specified. A broader toponym definition has consequences for the downstream tasks, such as corpora building and toponym recognition.

Gritta et al. classify toponyms (2020, pp. 691–692) in two major groups: *literal* and *associative*. The former refer to "where something is happening or is literally located" (ibid., pp. 691); this is the common definition of a toponym in the previous literature. Associative toponym is used "to modify *non-locational concepts* – – which are associated with locations rather than directly referring to their physical presence" (ibid., pp. 693). For example:

(1) We ate some Swedish [associative] meatballs in a pub in Paris [literal].

(1f) *Söimme ruotsalaisia lihapullia pubissa Pariisissa.*

The distinction here is that the literal toponym refers to a geographical location, while the associative one is linked to Sweden through the nationality, but the link to the physical location is slim. This distinction is evident through the *context* of the toponym: the study of how the interpretation of words and phrases changes in a context is called *pragmatism*, which is why Gritta et al. (2020) name their classification the pragmatic taxonomy of toponyms (ibid., pp. 685–686). They go beyond the top-level distinction and describe the various ways toponyms can be used as *modifiers* in phrases, for example:

(2) Our Tallinn hotel was really comfy.

(2f) *Tallinnan hotellimme oli hyvin mukava.*

Here, the toponym "Tallinn" modifies the *noun head*, which is "hotel". These types of phrases are often negligently overlooked according to Gritta el al. (2020, pp. 691–693), even though they carry geographical information as well. The taxonomy also includes demonyms (such as *French*) and

languages (*Russian*). Though these lack a location, they create ambiguity in at least English texts. The taxonomy cannot be described here in whole, but I recommend interest reader the relevant sections of the original paper (Gritta et al., 2020, pp. 690–694). However, for this thesis, Figure 2 shows the most salient results: it illustrates how a mere third (31.3 %) of toponyms in the GeoWebNews corpus were simple literal toponyms, the ones usually thought of as locations in geoparsing corpora. The main point of Gritta et al. (2020) is that geoparsers cannot be robustly used or evaluated if the central concept at the task's core, toponym, is not robustly defined.

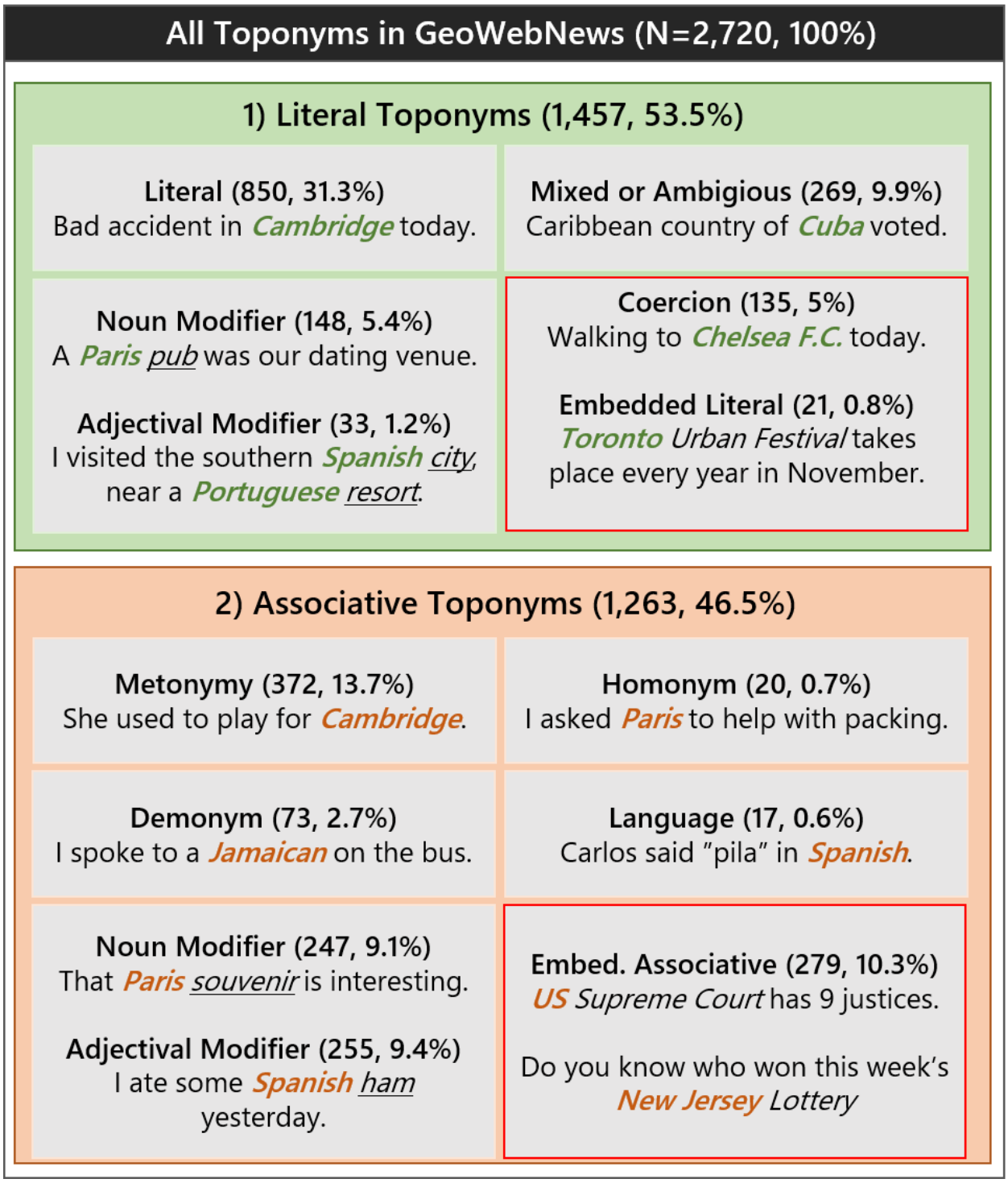


Figure 2. A taxonomy of toponyms recreated from Gritta et al. (2020, p. 690). The percentages show the share of toponyms falling into that class in the GeoWebNews corpus.

In general, there is scarcely agreement in previous research on which locations should be considered toponyms. This is evident when examining how toponyms are marked in corpora (plural of *corpus*).

Corpora are large collections of texts that have been annotated for some purpose: e.g., toponyms are annotated and linked with coordinates in geoparsing corpora to evaluate geoparsers (see Section 2.6 for more on geoparsing corpora). The only common feature type in most previous corpora are administrative units, like nations and cities. Natural features, roads and buildings are often not included, not to speak of loose toponym uses, such as metonymic use or demonyms (J. Wang & Hu, 2019b, pp. 10–11). There are bountiful edge-cases of toponyms that are hard to disambiguate, as Wallgrün et al. (2018) noticed when partly crowdsourcing the generation of their Twitter geoparsing corpus, *GeoCorpora*. These edge-cases (such as *United Nations* in a context where it could either be an organization or the UN headquarters) were hard for laypeople and experts in geography alike. Because it is tough to pin down one definition for toponyms in the literature, they are rather formed through the corpus-building process to best suite each use case (J. Wang & Hu, 2019b, pp. 10–11). Individual articles might similarly employ their own definitions based on their objective (see e.g. a Twitter use case by Gelernter & Mushegian, 2011, pp. 756–757).

2.3.2. Locations in time and space

The second part of geoparsing deals with linking toponyms with locations: finding a correct and suitable spatial representation for a platial concept. Neither toponyms nor locations are always static, clearly defined or even known. Places meaningful to people are bound to be named, but the names change due to political fluctuations, generational changes, memorable events occurring in that place, and numerous other reasons (Ainiala, 2018; Leidner, 2007, p. 71). The same place might hold different names over time, which Leidner (2007, p. 71) illustrates with the example of Karl-Marx-Stadt, better known historically and again today as Chemnitz. Historical, alternative forms and spellings (e.g. Helsinki/Hesa), and versions for different languages (Helsinki/Helsingfors) coexist for the same location (Purves et al., 2018, p. 214). There are also fictional places (such as *Hogwarts*), which cannot be located at all.

The locations change over time, which could be a consequence of cultural shifts or the physical world itself changing. For example, a municipality might gain ground in a merger or due to post-glacial rebound. The same toponym refers to locations in different time periods. Thus, an author writing about Lohja in the 1980s referred to vastly different area than in year 2021, since the municipality has gone through multiple mergers over time; features other than the area, such as population, were different as well. Or, as Goodchild and Hill (2008) put it, “places themselves come and go with the passage of time as well as the elements of place description” (pp. 1041). The

temporality of toponyms is a factor in geoparsing that should not be overlooked – for example, Tambuscio and Andrews (2021) found that toponym recognition proved difficult on historical texts from Armenia. Similar concerns made Bol (2013) call for a database where, e.g., the fluctuating administrative regions and their time periods would be stored: a "world-historical gazetteer" (ibid., p. 1089). There are in fact gazetteers exclusively for historical toponyms, such as the Pleiades (Barker et al., 2016).

Toponyms mark locations that are of wildly different scale, or granularity (Purves et al., 2018, p. 176). Think of a school building in the village of Koski in Hattula, Finland. The preceding sentence already marks four different locations, one a facility and the other three administrative areas of different levels. These form hierarchies: an explicit administrative hierarchy of municipality > province > country (Hattula > Kanta-Häme > Finland). Such relationships are often saved in gazetteers (Amitay et al., 2004) (see Section 2.5.1 for more on gazetteers). Spatial overlap, or containment, of the locations can also be used to express hierarchical relations: see Figure 3, where Koski is within Hattula, both are within Kanta-Häme and so forth. Also notice the varying sizes the locations are represented as. At the scale in Figure 3, the school building is not represented at all, Koski as point and the rest as polygons. As noted by Leidner and Lieberman (2011, p. 9), the varying granularity of the input poses a challenge for geoparsing.

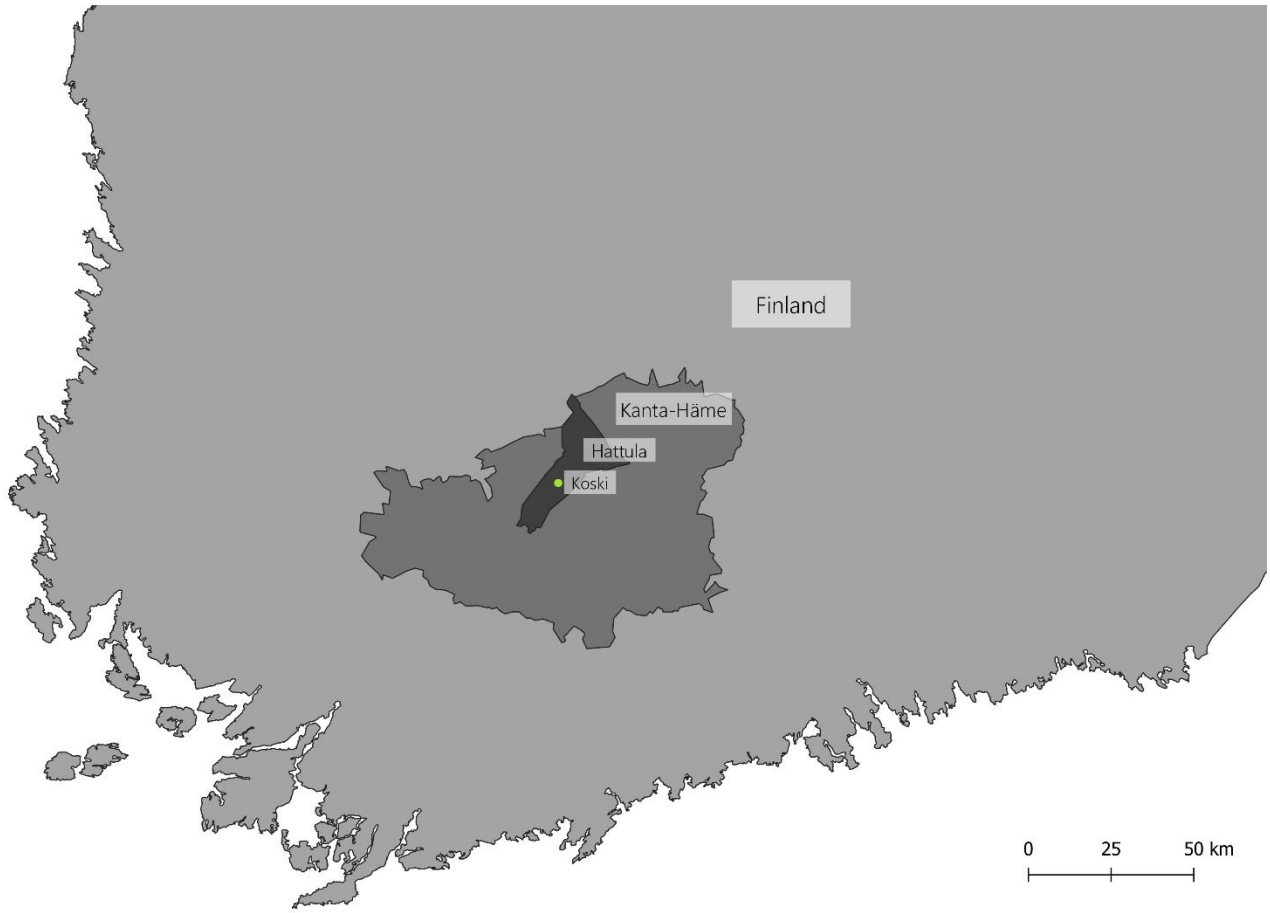


Figure 3. Visualizing containment and granularity differences in locations.

Not all toponyms define well-defined locations to begin with. *Vernacular toponyms*, such as *Piritori* in reference to the Vaasanpuistikko square in Helsinki, are used in everyday language but rarely codified in gazetteers, which mainly rely on official sources (Hu, 2018a, p. 82; Purves & Derungs, 2015). Recently, Hu et al. (2019) explored vernacular toponyms in a data-driven manner by finding and delineating local toponyms found in georeferenced online housing advertisements. Their linguistic-spatial method found a number of neighborhoods and points-of-interest not saved in traditional sources. A similar research topic are *vague cognitive regions*, areas that are used to mentally structure the environment. They are again rarely included in authoritative sources, and have vague extents and boundaries (Gao et al., 2017; Montello et al., 2014). An often studied example is Northern and Southern California in the United States, where their borders do not follow strict latitude lines and are more so influenced by the inhabitant's cultural expectations and perceptions (Montello et al., 2014).

Finally, there are cases where the location referred to linguistically is not actually the toponym (Leidner & Lieberman, 2011). For example, in "120 km North-East from Oslo", the toponym Oslo is simply used as a grounding landmark and the actual location is defined in the surrounding sentence. Efforts to automatically understand and formalize these types of complex geographical descriptions have been taken (see e.g. Du et al., 2017; Stock & Yousaf, 2018). Additionally, there have been calls to integrate geoparsers with the ability process these spatial descriptions (Gritta et al., 2017, pp. 621–622; Laparra & Bethard, 2020), but this line of research is still only budding.

2.4. Toponym recognition: Named Entity Recognition for geoparsing

2.4.1. Named Entity Recognition

The first step in geoparsing is the recognition of place names, or toponyms, in input texts. This task can be equated to a special case of *Named Entity Recognition* (NER) (Hu & Adams, 2021; Leidner & Lieberman, 2011), a widely-studied information retrieval task in the field of *Natural Language Processing* (NLP) (Ringland, 2016, p. 2). Named entities are anything that can be referred to by proper names and are deemed relevant classes for the task at hand. Traditionally, these classes include persons (*Aragorn*), locations (*Mirkwood*), organizations (*The Prancing Pony*) and geo-political entities (*Rohan*) (Jurafsky & Martin, 2022, pp. 164–165). In NER, the task is thus "to find spans of text that constitute proper names and tag the type of entity" (Jurafsky & Martin, 2022, p. 164). See Figure 4 for an example. Details like the number of entity classes vary by corpus – for example Luoma et al. (2020) include dates, products and events in their Finnish NER corpus, but have no separate class for geo-political entities – nations are grouped under locations instead. One span of text can usually only belong to a single class (*University of Barcelona* is tagged as an organization): unless a schema like *nested named entities* is used (*Barcelona* is a location within *University of Barcelona*) (Ringland, 2016). The task remains the same in toponym recognition, except that only locations are of interest.

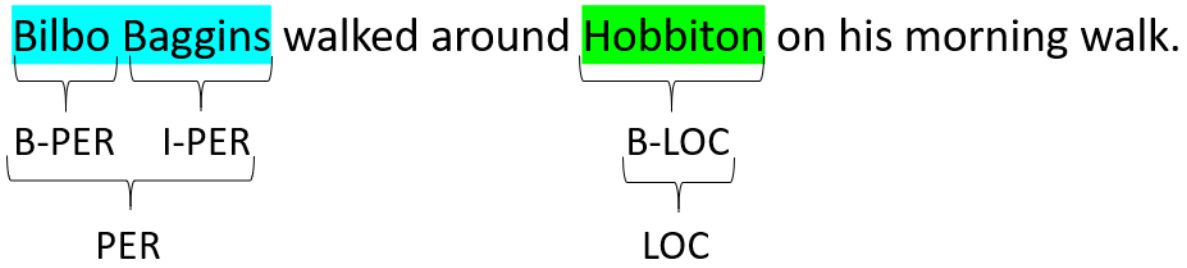


Figure 4. Example of a sentence with the named entities tagged (PER=person, LOC=location). The B and I markings follow the IOB2 tag format: B marks the beginning of an entity and I all the succeeding words belonging to the same entity.

2.4.2. Traditional approaches to toponym recognition

Toponym recognition is not a new problem. Leidner and Lieberman (2011) list three earlier approaches to the problem:

1. *Lookup*
2. *Rule-based*
3. *Machine learning based*

A simple approach would be to compare, or lookup, each word or character in text to a list of toponyms. This approach can be successfully applied if the study area and the target toponyms within are few and have little ambiguity (see Koivisto, 2021). However, for larger study areas ranging to global, ambiguities discussed previously (see Section 2.3) will quickly present problems: the same surface forms often have non-geographic meanings. For example, Dresden might refer to the German city (or one of many towns in the US), or it might refer to a film by the same name, or a jazz album by Jan Garbarek. The lookup could also fail if the surface form did not match the wordform in the lexicon (e.g. *Helsinki* in the conjugated form *Helsingissä*), though the surface form can be returned to a base form through computational means (*lemmatizing*), which alleviates this problem (Koivisto, 2021). In addition, the toponym list would inevitably be limited and need updating over time (Purves et al., 2018, p. 211).

The rule-based approach rests on the intuition that there are common linguistic features in the toponyms or their contexts which identify them. Therefore, a set of rules can be written in e.g. Regular Expression pattern matching language to exploit these. For example, to tag roads like *Maple Street*, a rule to match capitalized words followed by *Road*, *Street* or *Boulevard* can be constructed (Leidner

& Lieberman, 2011; Purves et al., 2018, p. 212). Rule-based geotaggers are quite common in geoparsing: for example the NER component of the Edinburgh geoparser applies manually crafted rules combined with lexicon lookup (Tobin et al., 2010) which performs quite well even in a recent evaluation (J. Wang & Hu, 2019a). As a downside, crafting the rules is time-consuming, language-dependent, requires expertise and it is still hard to account for cases such as colloquial language that does not follow the same patterns.

Toponym recognition can also be approached as a prediction task. A machine learning classifier is trained with examples (that is, in a supervised manner) to answer a question like: is this word a toponym given the input features? (Leidner & Lieberman, 2011; Purves et al., 2018, pp. 212–213). The input features have to be selected manually and can, for example, be preceding and succeeding part-of-speech tags (noun, verb etc.) or word affixes in the case of morphologically rich languages. A corpus of manually annotated gold-standard data, where the toponyms are known, is needed to train the classifier: the classification methods can be based on a number of options including Hidden Markov models or Conditional Random Fields (Purves et al., 2018, p. 213). Describing these models goes beyond the scope of this thesis, but an interested reader can find details and examples of their application to NLP tasks in Jurafsky and Martin (2022, Chapter 8).

The statistical machine learning approach has found success in toponym recognition, for example five of the six NER algorithms offered by the GeoTxt geoparser are machine learning based (Karimzadeh et al., 2019, p. 7). The downside of this approach is that it is data hungry, with the requirement of manually annotated training data. The classifiers might not generalize well to unseen datasets (Purves et al., 2018, p. 213), especially if the training data is from a different text domain. For example, a classifier trained on medical texts might do poorly on news articles.

2.4.3. Neural NER for toponym recognition

Recent advancements in toponym recognition are linked to the widespread introduction of *deep learning* methods in natural language processing, since these approaches often trump hand-crafted rules and the feature-based machine learning techniques in performance (Pilehvar & Camacho-Collados, 2020, p. 11). Especially effective are *contextualized word embeddings* and the Transformer-based language models that produce them (Pilehvar & Camacho-Collados, 2020, pp. 69–77). This topic is extensive and can be somewhat hard to grasp, but I will keep the explanations brief and mostly in the context of toponym recognition. To be clear on the terminology, an (artificial) neural network is a specific type of machine learning model that consists of layers of computing units (input, hidden

and output) that take in a value and output one to the next layer. Stacking multiple of these layers together creates a deep network, hence the term *deep learning* (Jurafsky & Martin, 2022, p. 133). In this thesis, I use the terms *neural* (*network*) and *deep learning* (*based*) interchangeably.

Word embeddings are a technique for approximating the meaning of words and are commonly used as input for neural NER models. Word embeddings are multi-dimensional vectors, basically lists of numbers, derived from word co-occurrence. The embeddings form a semantic space, where similar words are close to each other: for example, because *excellent* and *amazing* occur in similar contexts, they are close in the semantic space. (Jurafsky & Martin, 2022, pp. 106–107; Pilehvar & Camacho-Collados, 2020, p. 25). The embeddings can be static – *star* is represented by one vector whether its meaning in the sentence is a celestial object or a celebrity – or contextual, so that the embedding of the word is dynamic and dependent on the surrounding words (Pilehvar & Camacho-Collados, 2020, p. 77). BERT is a language model that can learn such contextual embeddings, trained on massive textual datasets (Devlin et al., 2018). Language models may be trained on texts from single (monolingual) or multiple languages (multilingual). The models form a base in which to ground different downstream tasks, such as NER. This process is called *transfer learning*: as a metaphor, think of first acquiring, through much hardship, general knowledge of Python, and then partaking a two-day class of Python-in-QGIS. The previously learned generalist skills carry over to the special task and make succeeding in it much easier. Similarly, a BERT model may be fine-tuned for tasks by training it with a small amount of labeled data (Devlin et al., 2018, p. 3; Jurafsky & Martin, 2022, Chapter 11).

These methods are relevant for this work because they have achieved state-of-the-art results in general NER. The high level of performance holds true for Finnish: Virtanen et al. (2019), show that their Finnish implementation of BERT with a NER layer achieves state-of-the-art results, beating a multilingual BERT model and a previous rule-based tagger (Virtanen et al., 2019, pp. 5–6). Does the same level of high performance carry to toponym recognition? J. Wang and Hu (2019a) report on a recent geoparsing competition and state that all the winning teams used neural network based models in their geoparsers. They further compared the competition geoparsers to older systems and run them against test corpora. The results show that these geoparsers, such as the winning DM_NLP (X. Wang et al., 2019), perform top of the line on most, but not all, of the eight datasets. Recently, a neural network toponym recognizer built for social media texts was introduced, and once again achieved commendable results (J. Wang et al., 2020).

The results from multiple fronts give confidence that the methods based on deep learning architectures are good option and perhaps the direction to go for toponym recognition. A question remains: why do these models perform so well? Central to this thesis is the deep language model BERT. Though BERT is still fresh and somewhat opaque technology, Rogers et al. (2020) note in their review several features of it that could explain its performance. BERT embeddings have syntactic, semantic and, to a limited extent, general world knowledge encoded in them. I presume these features help alleviate geo/non-geo ambiguity. It is also not stumped by out-of-vocabulary words (words the model did not encounter when training) in the input as easily as traditional machine learning models – this is because the embeddings are actually sub-word tokens (Jurafsky & Martin, 2022, pp. 246–247). This means that if that if a whole word is not found in the vocabulary, the program splits it to smaller parts, each with their own embeddings: for example *koulussako* → [koulu, ##ssa, ##ko].

2.5. Toponym resolution

Toponym resolution (also known as *geocoding*) is the second task of a geoparsing process. In short, once toponyms in the input have been recognized, they must be unambiguously tied to some location on Earth. To do so requires both knowledge of *where* the toponyms should be located and *which* is the correct interpretation in case of geo/geo-ambiguity.

2.5.1. Gazetteers

Gazetteers are databases of toponyms, the toponyms' spatial footprints and various attribute information tied to the toponyms (Hill, 2006, pp. 91–92). For example, the city of Espoo has an entry in the GeoNames gazetteer with the spatial footprint in the form a coordinate pair (60.25, 24.667), a unique GeoNames ID (660158) and population as additional information. In geoparsers, gazetteers are used to get coordinates for the identified toponyms and often to disambiguate between candidate toponyms (Purves et al., 2018, pp. 215–217). While some geoparsing approaches function independently of gazetteers (e.g. DeLozier et al., 2015; Hulden et al., 2015), most are database-dependent in training or while running them (see e.g. systems explored by Gritta et al., 2017; J. Wang & Hu, 2019b), which merits the closer examination of gazetteers.

Table 1 lists several prominent open-access gazetteers, particularly ones that are used by geoparsers. To represent Finnish placename databases, a dataset by the National Land Survey of

Finland (*Maanmittauslaitos*) is included; the other gazetteers have global coverage. The gazetteers vary by the sources of information, with some derived from authoritative sources, OpenStreetMap being crowdsourced and GeoNames augmenting authoritative data with user input (Acheson et al., 2017). OpenStreetMap, which primary purpose is to be an online map and not a gazetteer, contains points, lines and polygons as spatial features, while the rest only contain coordinate points. Easily the most often used database for geoparsing is the aforementioned GeoNames[2] used as the only or additional data source by for example the Edinburgh geoparser (Tobin et al., 2010), DM_NLP (X. Wang et al., 2019) and GeoTxt (Karimzadeh et al., 2019).

Table 1. Examples of publicly available gazetteers

Gazetteer	**Coverage**	**Footprint**	**Data source**
GeoNames (GN)	Global	Points	Authoritative / volunteered
Getty Thesaurus of Geographic Names (TGN)	Global	Points	Authoritative
NLS Geographic names	Finland	Points	Authoritative
OpenStreetMap	Global	Points, lines, polygons	Volunteered

Gazetteers vary both in their stated purpose and in quality on different metrics – how current their data is, how accurate it is, what spatial scale it covers and how uniformly the features are spread spatially and across feature types (Hill, 2006, p. 107). Acheson et al. (2017) explored these properties quantitatively in GeoNames and TGN, and found that the gazetteers do not cover the globe equally. For example, there are strict discontinuities in feature densities in national borders, such as hills in Norway being exceptionally well mapped in comparison to Finland. The authors note that especially less developed countries may be poorly covered by these global gazetteers, which skews results when using them for tasks such as toponym resolution (Acheson et al., 2017).

[2] https://www.geonames.org/

The production and collection process of place names has also been critically examined: whose description of space is recorded? (Rose-Redwood et al., 2010) For example, the collection practices in Finland used to promote collecting place names from elderly males (Ainiala, 2018). These imbalances are bound to flow downstream to the gazetteers, which often draw from authoritative sources (Table 1). In the same way maps do not merely convey information about the world objectively and neutrally (Harley, 1989), gazetteers are not arbiters of absolute truths about our world. They cannot describe the world fully, the spatial experience of everyone, or solve disputed localities: Cope and Kelso (2015) see gazetteers as "the space where debate about place is *managed* but not decided". While keeping these ponderings in mind, gazetteers are nonetheless a necessary tool in geoparsing.

2.5.2. Toponym resolution methods

Correct and suitable geographical representations must be produced for the toponyms after they have been recognized. To this end, gazetteers may be queried: the candidates that match the input string are returned. When there are multiple candidates for a single toponym, they must be disambiguated using some type of method. Similarly to toponym recognition and NER, this task is not unlike the general *Named Entity Disambiguation* task (Santos et al., 2015). However, I will focus on specifically methods proposed for grounding place names.

Buscaldi and Rosso (2008) group these methods into three major categories:

1. *Map-based*, which use the spatial analysis methods.
2. *Knowledge-based*, which draw from external resources, like gazetteers to rank the candidates.
3. *Data-driven* or *machine learning based*, where a classifier is trained often based on the surrounding context of the toponym.

Please note that, although I present these approaches individually, toponym resolution systems often leverage multiple methods, all of which do not fit neatly to Buscaldi's and Rosso's (2008) scheme.

The most prominent spatial disambiguation method is *spatial minimality*, as introduced by Leidner (2007). It relies on the assumption that places mentioned in a single document tend to be close to each other, and thus finding a solution that minimizes the distance between the resolved toponyms would be the correct one. For example, let us examine this sentence:

(3) New school is being built in Koski, Hattula
(3f) Hattulan Koskeen rakennetaan uutta koulua

Let us say Hattula is unambiguously resolved to the municipality in Kanta-Häme. Following the principle of spatial minimality, the village of Koski that is near Hattula will be selected because it leads to the smallest minimum bounding box (Leidner, 2007, pp. 148–153). The assumption that toponyms in the same document are most likely spatially near might not hold for all text types: for example, spatial minimality provided no performance improvements when run on the GeoCorpora corpus of tweets (Karimzadeh et al., 2019).

What Buscaldi and Rosso (2008) call knowledge-based methods exploit external sources to get ranking information. Often this ranking aims at selecting the most prominent location, be it through population count, type categories, the number of alternate names or some combination of the aforementioned. Selecting the candidate with the highest population is a common heuristic, and one that provides robust results; for example functioning as a baseline method (J. Wang & Hu, 2019a) and providing a performance boost of 50 percentage points over simple string matching in GeoTxt (Karimzadeh et al., 2019). Population heuristic is also easy to implement, since such information is readily stored in gazetteers; however, it of course only works for populated places.

Disambiguating may also be done with place types, such as preferring cities over water features, or using the number of alternate names as a proxy of a feature's prominence (Karimzadeh et al., 2019). Hierarchical containment relationships may also be exploited (Purves et al., 2018, p. 217, also see Figure 3): for example in (3), Koski is a subsection of Hattula and both of them are subsections of Kanta-Häme in the administrative hierarchy. Finally, specific rules might be crafted if they suit the use case: for example, Alex et al. (2015) report exploiting the nearness to 19th century ports as a feature that aided in resolving toponyms in historical trading documents.

Data-driven models use the surrounding context of the toponyms as features to predict which location is the correct one. Businesses, landmarks, celebrities, dialectal terms and so forth have a location on Earth after all, and the intuition here is that these contextual tidbits can be used to determine the location referred (Ju et al., 2016; Santos et al., 2015). Data-driven models can function independently of gazetteers by modelling the inherent geoindicativeness of words (DeLozier et al., 2015; Hulden et al., 2015; Ju et al., 2016). As an example system, Hulden et al. (2015) divide the globe into a grid and associate individual words in georeferenced tweets with the grid cell they fall in. Thus, e.g. regional Spanish dialectal words get associated with cells in Mexico, Argentina etc. (Hulden et al., 2015, p. 146). Predicting the location of new documents is tasked to probabilistic classifiers, which use the distribution of words and documents as training features. In another approach, Ju et al. (2016) combine knowledge-driven approach of finding entity co-occurrences in

Wikipedia texts and DBpedia database queries with topic modelling geo-referenced Wikipedia articles. Similarity between the topic-models and new toponyms and their contexts are measured to disambiguate the toponyms.

The latest data-driven toponym resolution models benefit from the recent advances in deep learning (Cardoso et al., 2019; Gritta et al., 2018; Kulkarni et al., 2020), mirroring the strides made in toponym recognition. Similarly to the named entity recognizers that use word embeddings to better represent language and identify the desired named entities, these new systems encode linguistic and sometimes geographic knowledge in feature vectors. For example, CamCoder (Gritta et al., 2018) creates feature vectors of the toponym itself, all the surrounding toponyms and context of up to 200 words on either side of the toponyms. Feature vectors from the textual context are created elsewhere with ELMo (Cardoso et al., 2019) and GloVe (Kulkarni et al., 2020). Gritta et al. (2018) also create a geographical feature vector, named MapVec. MapVec is built by embedding the coordinate point location of each candidate for a toponym, biasing with population, which creates a probability surface on a global grid. The grid is then collapsed into a 1-dimensional vector. These features are analyzed by multiple layers in the network, until the four vectors are concatenated and the combined information is used to predict the correct location.

The latest deep learning toponym resolvers perform the best in the tests presented in the papers (Gritta et al., 2018; Kulkarni et al., 2020), though this difference is not clear-cut. Even simple population heuristic is a commendable baseline to beat: it is relatively simple to resolve something like *Paris* to the most probable referent. The task's complexity increases significantly when trying to correctly resolve the few cases of *Paris* that are referring to the city in Texas. For those cases, the deep learning resolvers could prove to be essential.

2.6. Geoparsing evaluation

How do we know whether geoparsers work as intended? Their performance must be evaluated in some manner. To achieve this, researchers have produced corpora (singular: corpus), which are collections of texts that have been annotated with labels. In geoparsing corpora, the toponym spans (*New York*) are marked and each toponym gets attached with location information, which is almost always a coordinate pair or an identifier in a gazetteer. The corpora cover a range of text genres, or domains, such as *Local-Global Lexicon* (Lieberman et al., 2010) and *GeoVirus* (Gritta et al., 2018) for news articles, *GeoCorpora* (Wallgrün et al., 2018) and Matsuda et al. (2015) for tweets and *WOTR* (DeLozier et al., 2016) for historical documents. The annotations in the corpora are taken as ground

truth and compared to what the geoparsers output: the closer geoparsed result is to the ground truth, the better. As with geoparsing in general, the evaluation should also be two-pronged: toponym recognition and resolution should be evaluated with different metrics (Karimzadeh, 2016). While there has been significant variation in the measurements used, recent research has proposed and standardized evaluation metrics (Gritta et al., 2020; J. Wang & Hu, 2019b), which I present below.

2.6.1. Toponym recognition evaluation

As mentioned in Section 2.4.1, toponym recognition can be thought of as a subtask of named-entity recognition. The performance metrics for NER are well-established: precision, recall and F_1 score (Jurafsky & Martin, 2022, p. 178; Ringland, 2016, p. 190), sometimes appended with accuracy. The same metrics are adopted in toponym recognition (see e.g. DeLozier et al., 2016; Gritta et al., 2020), most prominently in the recently introduced geoparser evaluation platform *EUPEG* (J. Wang & Hu, 2019b). For all these metrics, the geoparser is run on the texts. The output is then compared to the gold-standard annotations: to see if the system marks the same words or spans of text as the human annotators have. Let us briefly explore the measures (presented in e.g. Jurafsky & Martin, 2022, pp. 67–69; J. Wang & Hu, 2019b):

- **Precision** is the ratio of correctly recognized toponyms (called *true positives*) against all toponyms labeled by the system, including falsely labeled ones (called *false positives*). A high number of falsely labeled entities will worsen the score: scores closer to 1 are better. If true positive is TP and false positive is FP, then precision is:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall** is the ratio of true positives against all toponyms in the corpus. A high number of missed toponyms (called *false negatives*) will worsen the score. If true positive is TP and false negative is FN, then recall is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F_1 score** or **measure** is the harmonic mean of precision and recall. The score is negatively affected if either of the measures are low (J. Wang & Hu, 2019b, p. 14).

$$\mathrm{F}_1 \text{ score} = 2 * \frac{Precision\ x\ Recall}{Precision + Recall}$$

Let us further elaborate on the metrics through an example. Let the input sentence be:

(4) "Lohja Inc. has offices in Uusikaupunki, Espoo and Helsinki", said Aino Lahti, the CEO

(4f) "*Lohja Oyj:n toimistot sijaitsevat Uudessakaupungissa, Espoossa ja Helsingissä", sanoi tj Aino Lahti*

Figure 5 shows an imaginary toponym recognition system applied on sentence (4). The system correctly labels Espoo and Helsinki as locations, but then incorrectly marks a corporation (Lohja Inc.) and a person (Lahti) as locations, which are false positives and misses Uusikaupunki, which is a false negative. Placing these results in the formulas above, we acquire precision of $\frac{1}{2}$, recall of $\frac{2}{3}$ and F_1 score of roughly 0.57.

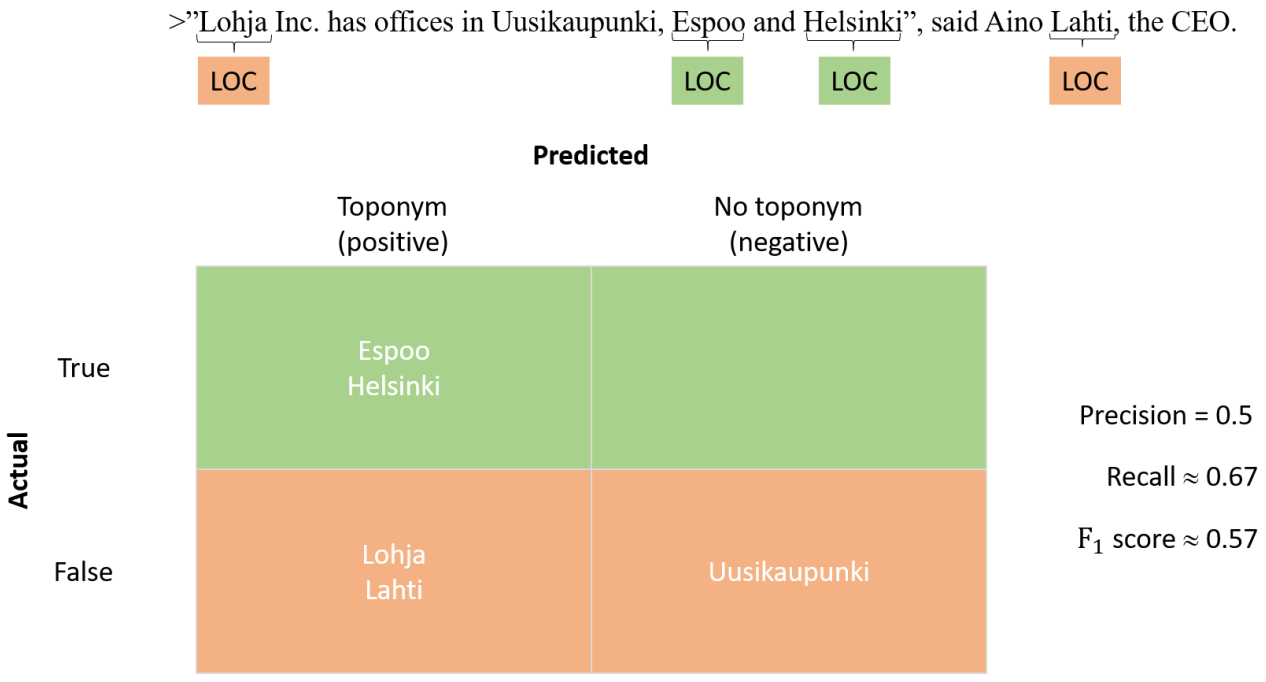


Figure 5. Example of toponym recognition evaluation. The sentence above has been analyzed by an imaginary NER system; the annotations it has given are below the sentence. The results are placed in a confusion matrix and the evaluation metrics reported beside it.

2.6.2. Toponym resolution evaluation

The second evaluation section aims at determining whether the toponym was correctly located. In the case that the system returns a feature ID, such as GeoNames ID 658225 for Helsinki, and the same ID is included in the corpus, evaluation can be done with a simple accuracy score (Karimzadeh, 2016). This approach fails if no ID is saved in the ground-truth data, and moreover, it is dependent on the gazetteer used. If the geoparser uses OSM as a source gazetteer, but the corpus has GeoNames IDs, the evaluation cannot be done.

That is why more common approaches evaluate the Euclidian distance from the predicted coordinate point to the ground-truth one: this is called *Error Distance* (ED). If the ED is 0, that is to say, if the predicted coordinate and the ground-truth one are identical, the toponym is completely correctly resolved. If the error distance is > 0, the magnitude of the ED hints at the significance of the error (for example, ED of 1000 km is probably more significant than ED of 10 km).

There is yet no agreement on what is the best approach at evaluating toponym resolution (Gritta et al., 2017, p. 612), rather, each metric highlights some attribute of the geoparser's performance. I will introduce the four approaches adopted in the EUPEG platform (J. Wang & Hu, 2019b). Please note that, for the sake of space and the emphasis of this thesis, I will attempt to explain only the intuition behind the metrics. For the exact formulas, the reader is referred to e.g. J. Wang & Hu (2019b, pp. 15–17).

- **Mean Error Distance (MED)** is the mean distance between the ground-truth coordinates and the predicted ones in kilometers. A simple metric to distill the average error, however, it is sensitive to extreme outliers since the errors are not usually normally distributed. In geoparsing, most errors are relatively small, but the error distances for the fifth quintile are very large (see Gritta et al., 2017, p. 613).
- **Median Error Distance (MdnED)** is the same as MED but with the statistical value of median. Not as prone to outliers, but perhaps no better at representing the error distribution.
- **Accuracy@k** is the percentage of predictions located within *k* distance of their respective ground-truth coordinates. As a convention, *k* has been set to 161 kilometers (equaling a 100 miles), which is why the metric is often referred to accuracy@161. The intuition behind this metric is that it can handle cases where the toponym has been correctly disambiguated, but the location is different for the geoparsed one and the ground-truth one (due to for example relying on different gazetteers) (J. Wang & Hu, 2019b, p. 14). A weakness of accuracy@k is that it is not sensitive to the magnitude of error as long as the location falls within *k*: for

example, a distance of 100 km from the ground-truth is treated the same as a distance of 10 km, when k=161 (Gritta et al., 2017, p. 612).

- **Area Under the Curve (AUC)**, introduced for this purpose by Jurgens et al. (2015), is an attempt to address the shortcomings of the previous metrics. Once again using the error distances but this time sorting them in a rising order and transforming them by taking the natural logarithm of the error distances. The upper bound of the error is the maximum distance between two points on Earth, approximately 20,038 km. The area under the curve is the proportion of the total area of the plot that these errors cover: the returned value falls between [0,1] and smaller is better. AUC follows the intuition that a small difference in a small error (e.g. 10 km vs 30 km) is more significant than a small difference in a large error (e.g. 510 km vs 530 km). That is to say, under AUC, the errors do not scale linearly, and the metric punishes geoparsers for small errors more harshly than large ones. Figure 6 shows this visually and an applied example is presented in Figure 12.

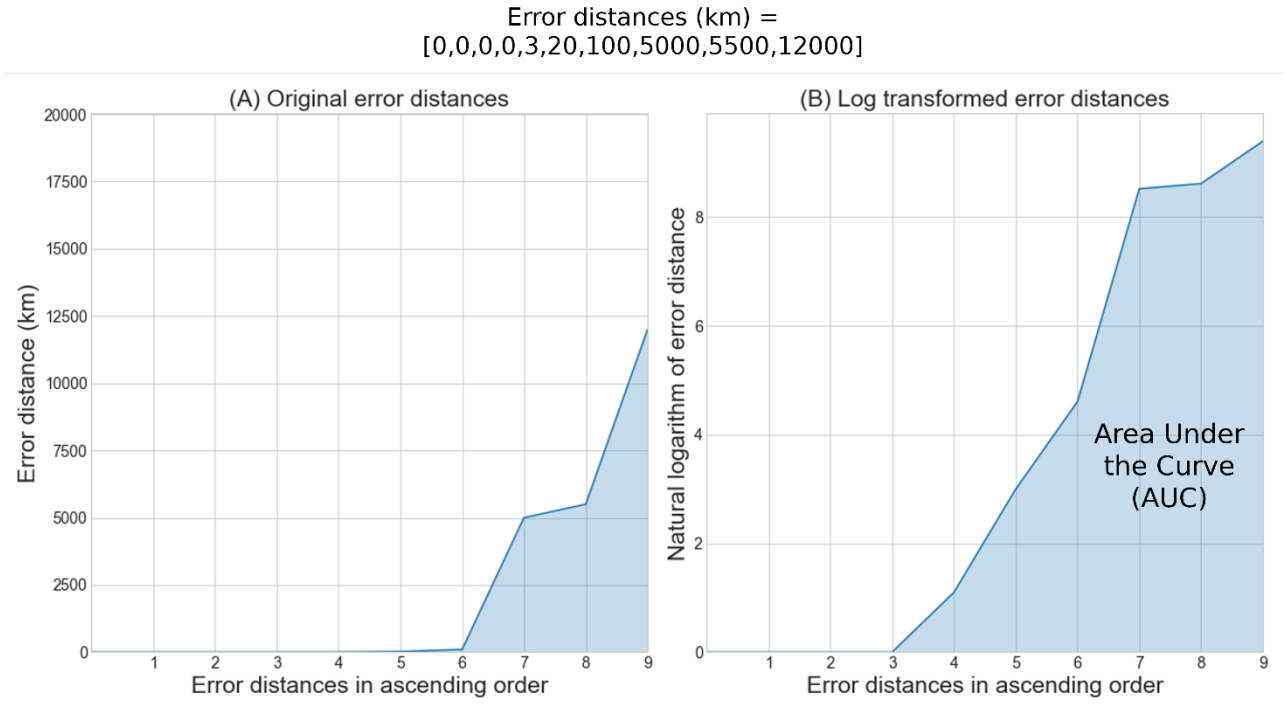


Figure 6. AUC visualized with ten error distances plotted, after Gritta et al. (2020, p. 696). (A) showcases error distances plotted as they are and (B) after ln transformation. The error distance data points are listed at the top of the figure. The area covered by the blue line is the "area under the curve", and the value AUC is calculated by measuring its share of the total area (the maximum possible errors). The upper bound of the Y axes are max error distance, or half the Earth's circumference. If all the points were correctly located, AUC would be 0 and if they would be furthest possible, AUC would be 1. In (B), the AUC value is 0.342.

Let us again elaborate through an example: let us apply an imaginary toponym resolver on the three toponyms in sentence (4). Espoo is resolved correctly, but it seems Uusikaupunki and Helsinki are off by 35 and 200 km respectively. These are the error distances, which are used to calculate the performance metrics. The error distances and the metrics are presented visually in Figure 7.

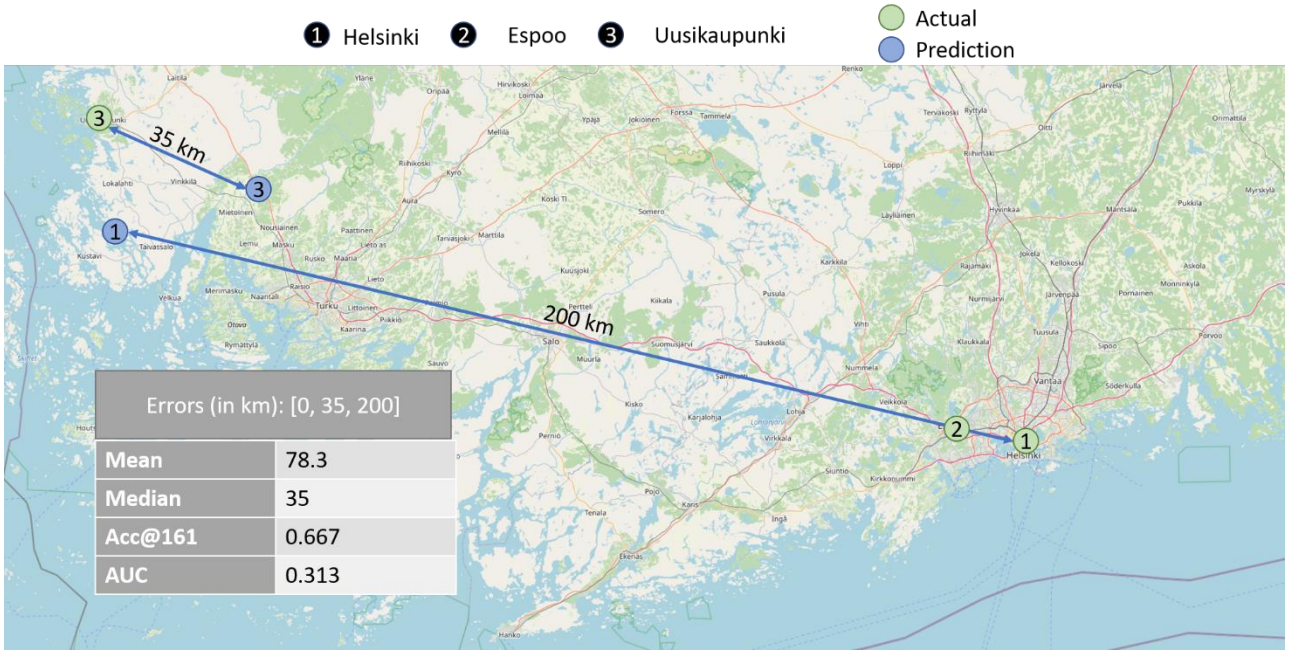


Figure 7. Example of toponym resolution errors and error metrics. Background map © OpenStreetMap contributors.

2.6.3. Factors affecting geoparser performance

What explains performance variation of geoparsers? Naturally, the toponym recognition and resolution methods affect performance greatly; these are covered in Sections 2.4–2.5. Another important factor are the corpora and the choices made when building them i.e. what text domains are included and which entities are annotated within. The effect of these choices are reflected in the vastly varying performance of geoparsers, as seen in recent evaluations (Gritta et al., 2017; J. Wang & Hu, 2019a).

Texts differ in how formal they are: for example, news articles are expected to have consistent spelling and capitalization, and use standard vocabulary. Social media posts, such as tweets, may in turn be highly informal, containing emojis, novel spellings and words. Not to mention tweets are

short, which limits their context (Carter et al., 2013). Indeed, it seems geoparsers perform better on GeoVirus (Gritta et al., 2018), a corpus of disease news articles, than on GeoCorpora (Wallgrün et al., 2018), which contains tweets (J. Wang & Hu, 2019a).

Another axis relates to how ambiguous and fine-grained toponyms are contained in the corpora. Many of them purposefully contain ambiguous toponyms to test geoparser performance under challenging circumstances. The Local-Global Lexicon contains articles from local newspapers. The articles frequently refer to local toponyms that share names with global cities (London, Ohio vs. London, UK) (Lieberman et al., 2010). Similarly, Ju et al. (2016) built a corpus from web queries and purposefully selected ambiguous sentences from the returned websites. Both corpora may be challenging especially for simple population heuristics (Section 2.5.2) and indeed, many geoparsers struggle with them (J. Wang & Hu, 2019a).

The toponym definition matter as well: many corpora only include nations and cities, while others have landmarks and streets annotated too (J. Wang & Hu, 2019b, p. 11). Fine-grained toponyms are of course harder to geoparse, but this variation has other effects as well. If a geoparser is trained to find buildings but buildings are not annotated in a corpus, the geoparser's precision gets worse when it annotates them.

3. Data and Methods

In this chapter, I describe the development of Finger (**FIN**nish **GE**oparse**R**) and creation of the corpora used to evaluate its performance. In other words, the workload of this thesis can be divided into three tasks:

(1) Creating the geoparser
(2) Creating evaluation data
(3) Evaluating the geoparser

The workflow of each task and how they connect to each other are presented in Figure 8. The workflow of the thesis. See Table 2 for a description of the input datasets. The creation processes are elaborated on in the following sections (3.1–3.3). All input datasets are collectively introduced in Table 2 and will similarly be expanded upon in the following sections.

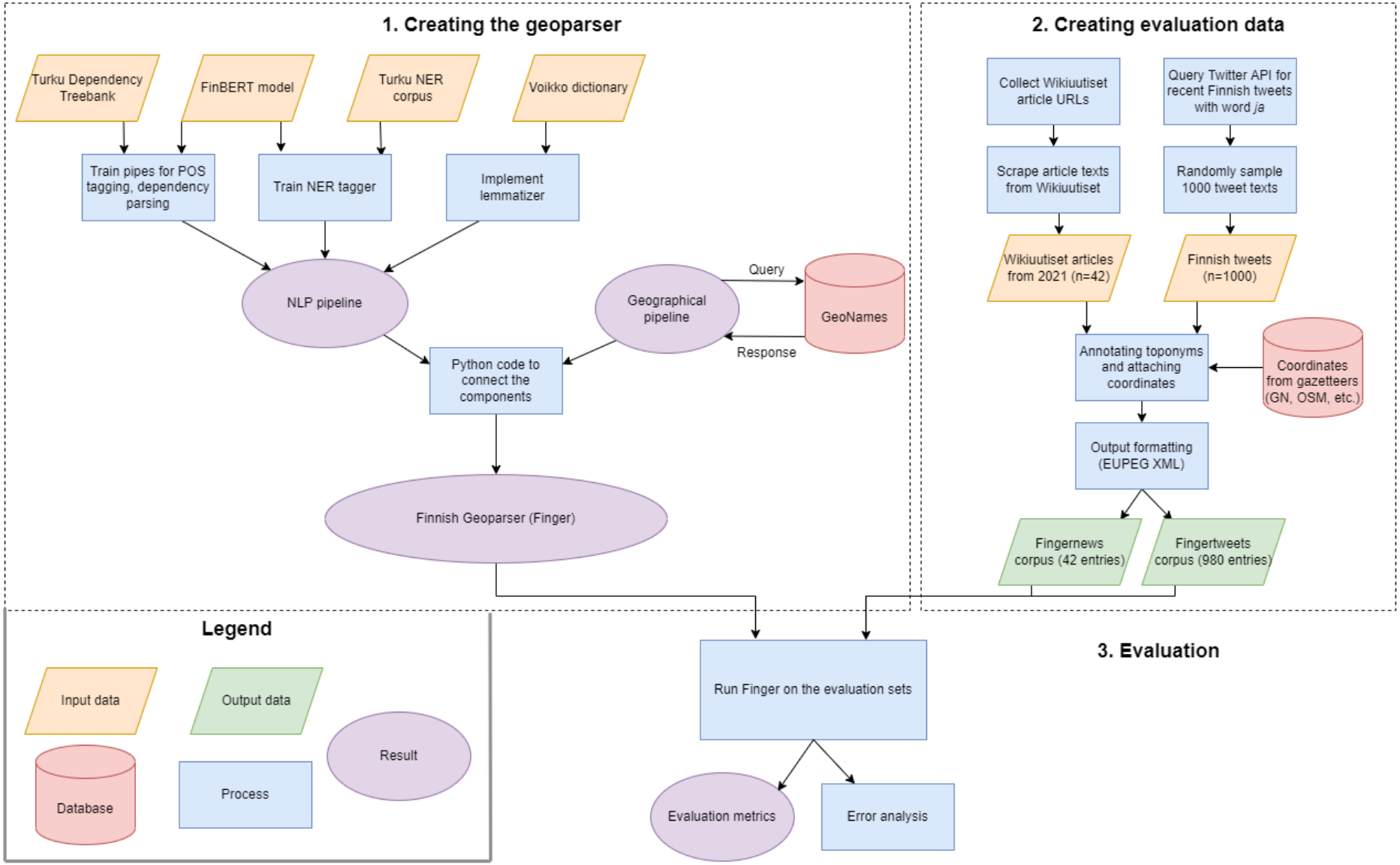


Figure 8. The workflow of the thesis. See Table 2 for a description of the input datasets.

Table 2. Input datasets and their purpose in this study

Dataset	Description	Purpose	Source
FinBERT model	A large pre-trained language model that offers state-of-the-art performance for Finnish NLP tasks.	The model is fine-tuned for the tasks of NLP processing and NER tagging using the respective corpora.	(Virtanen et al., 2019)
Finnish tweet set	Finnish language tweets containing the word *ja* queried from Twitter's API and randomly sampled to 1000.	To annotate and create evaluation material: Finger-tweets.	Gathered by the author
Turku Dependency Treebank (TDT)	A large corpus tagged with dependency graphs, part-of-speech tags, lemmas and other linguistic tags according to the Universal Dependencies project.	Future-proof the geoparser by training the NLP pipeline for general NLP tasks.	(Haverinen et al., 2014)
Turku NER corpus	Named Entity Recognition corpus consisting of texts from TDT that have been marked with a set of entity tags. Importantly, including location.	Train the general NER tagger, which is then used for toponym recognition in Finger.	(Luoma et al., 2020)
Voikko dictionary (Joukahainen)	Dictionary covering roughly 40,000 Finnish words and conjugation information.	Enables lemmatizing: the lemmatized wordform is looked up in this dictionary.	[3]
Wikiuutiset articles	All articles for the year 2011 from the Finnish Wikinews.	To annotate and create evaluation material: Finger-news.	Gathered by the author

[3] https://www.puimula.org/htp/testing/voikko-snapshot-v5/

3.1. Creating Finger

The aim of this thesis is to create a general-purpose open-source Finnish geoparser. By general purpose, I mean that the geoparser is not custom-built for any particular text type: the (perhaps lofty) aim is that it can effectively process anything from informal blog posts to legal documents. Bearing in mind the limitations of a master's thesis project, Finger should be informed by the latest developments in English geoparsing research and should be easily applicable for research that benefits from georeferenced data. Another guiding principle was to use existing resources (tools, datasets) whenever possible. Finger's source code and installation instructions are shared in a GitHub repository[4] under MIT license.

As described in Chapter 2, geoparsing consists of toponym recognition and resolution. I similarly decided abstract these tasks to two pipelines: *Natural Language Processing* (NLP) *pipeline* and *geoinformatics* (GIS) *pipeline*. A pipeline consists of individual *pipes*, or parts that run a certain section of the analysis and feeds its output forward. Therefore, the NLP pipeline recognizes toponyms and handles similar linguistic tasks; the GIS pipeline resolves the toponyms to locations; finally, a structured output is returned to the user in a format of their choosing. Finger, written in Python, wraps these pipelines and runs the geoparsing process. Finger is a *toponym geoparser*, that is, it does not try to locate whole documents. Instead, it simply attempts to geoparse each toponym it recognizes and leaves any further analyses to the user. Finger is suitable for batch processing which means that it accepts, and is built for, multiple inputs. An example of Finger in operation is given in Figure 9. Four example sentences and how the system handles them are shown. The example cases are:

(1) The input contains a toponym and it is correctly resolved.
(2) No toponyms.
(3) A toponym is recognized and lemmatized, but resolving it fails.
(4) Two toponyms: the first is incorrectly lemmatized and not resolved. The second is correctly geoparsed. A tuple with these results is returned.

[4] https://github.com/Tadusko/fi-geoparser

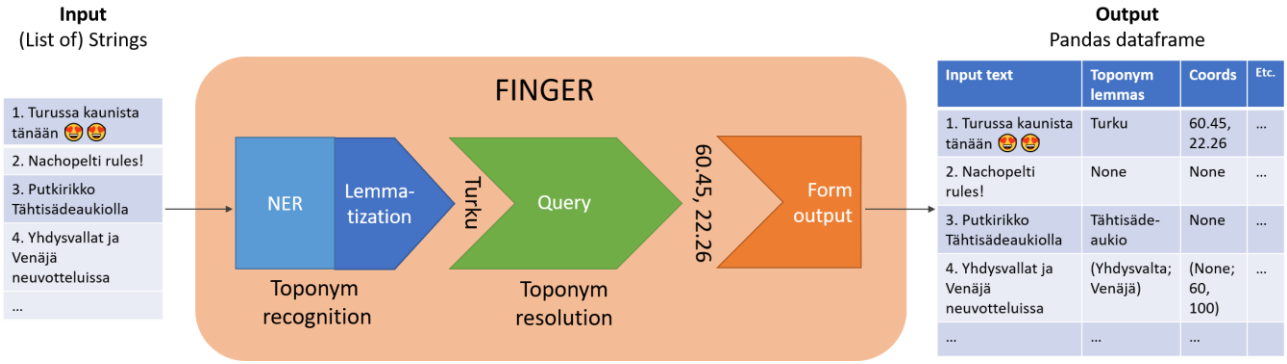


Figure 9. How Finger operates, with the most crucial components highlighted. The example dataframe differs from the one actually output by Finger: it has been truncated and some columns have been renamed for clarity.

3.1.1. NLP Pipeline

NLP pipeline is the more extensive of the two pipelines. It consists of three pipes:

(1) General NLP pipe
(2) NER tagger
(3) Lemmatizer

All of the pipes were trained and joined to a pipeline using the spaCy natural language processing library for Python (Honnibal et al., 2020). The NLP pipeline, about 1 GB in size, is simply installed to a Python environment with spaCy, after which it is usable in Finger. The pipeline can be easily updated and reinstalled in the future. A pre-trained Finnish BERT was recently published by Virtanen et al. (2019). I use it in pipes (1) and (2) as the base model that is fine-tuned for Universal Dependencies and NER tagging respectively. The training data for the Finnish BERT was gathered from three sources: news articles, online discussion (from Suomi24 forum) and an Internet crawl. After data cleanup, these sources amount to ≈21 Million documents and ≈3.3 Billion tokens (words separated by whitespace).

The first pipe, the general NLP pipe structures the raw input text according to the Universal Dependencies (UD) annotation schema. Universal Dependencies provides a framework for the computational analysis of various linguistic features (de Marneffe et al., 2021; Nivre et al., 2016). Relevant for this work are linguistic tags, such as parts-of-speech (for example: noun, verb and adjective) and morphological features (for example: is the word in singular or plural form). A tagger to structure text with these features can be trained with a Finnish UD corpus: the corpus used in this work was created by Haverinen et al. (2014). The general NLP pipe outputs are not currently used in

the latter pipes: it is included as a way of futureproofing should a more comprehensive linguistic understanding help in toponym recognition or resolution.

Fine-tuning the BERT model for named entity classification requires a NER corpus. At the time the work was done, there were two Finnish NER corpora, one based on technology news (Ruokolainen et al., 2020) and another covering a wider range of text domains from speeches and blogs to Wikipedia and finance articles (Luoma et al., 2020). Because I expect the geoparser to handle all types of input texts, I chose to use the broad-scope corpus by Luoma et al. (2020) in this work. Six entity classes are annotated in the corpus: of those, locations are crucial for this work. Annotation schema of the corpus (for example, the definition of a location in the corpus) affects all downstream tasks. That is why some annotation decisions are worth highlighting. Unlike some NER corpora, this corpus does not include a separate tag for political and administrative entities (nations, provinces etc.), which are all grouped under locations. The definition of location[5] thus includes administrative entities as well as natural features (rivers, mountains), buildings, roads and even astronomical features (planets and stars). As a further motivation for selecting a deep learning based method in this work, a NER implementation based on Finnish BERT performed the best out of four tested methods on the corpora's internal test set (Luoma et al., 2020, pp. 4620–4621),

Corpora used in training machine learning classifiers are often divided to distinct sets, namely development, training, and testing (dev, train, test). Training set contains the examples provided to the model during training. Dev set is used to measure progress during training. Test set is used for evaluation: those examples are used to measure how well the model generalizes, that is, performs on examples it has not encountered before. I fine-tuned the model using spaCy and the default hyperparameters (see the training files here[6]). Describing the fine-tuning process in detail goes beyond the scope of this thesis, but details can be found in Jurafsky and Martin (2022, Chapter 11). Basically, the language model (Finnish BERT, in this case) is tasked with predicting the correct named entity class by feeding it training examples from the NER corpus.

Results from the test section run are reported in Table 3. The performance is a few points lower than the results reported by Luoma et al. (2020, pp. 4620–4621), especially for the location tag, where precision is almost 10 points worse. The difference is probably, at least partly, explained by the selection of hyperparameters, which are the settings governing the training process. A good choice

[5] https://github.com/TurkuNLP/turku-ner-corpus/blob/master/docs/Turku-NER-guidelines-v1.pdf, pp. 7-9
[6] https://github.com/Tadusko/finger-NLP-resources

of hyperparameters, so called hyperparameter tuning, could improve the performance, but was not done in this work.

Table 3. Performance of the NER classifier trained in this work on the internal test set. Overall covers the performance across all entity tags and Location only to entities tagged as location.

Entity type	**Precision**	**Recall**	**F-score**
Overall	0.8783	0.8890	0.8836
Location	0.8418	0.9236	0.8808

This NER classifier is used as-is in Finger. The words classified to the location class are collected by Finger and passed on to the lemmatizer. After NER, Finger has a functionality to filter the recognized toponyms according to handcrafted rules: currently, the only filter is that the toponyms must be longer than one character long. This filters out most emojis, which were a common source of error (see Section 4.2)

The lemmatizer is the final component of the NLP pipe. In lemmatization, the varying surface forms toponyms can take in Finnish (*Oulussa, Ouluun, Oulunhan* etc.) are attempted to return to a base form, also known as lemma (*Oulu*). Finger uses a simple lemmatizer provided by the NLP toolbox Voikko[7]. Voikko looks up words like *Oulussa* in an open-source dictionary called Joukahainen[8]. Joukahainen morpho, the dictionary used in this work, includes about 40,000 entries and the conjugation paradigms related to them. Using the conjugation paradigms saved alongside the dictionary entries, Voikko attempts to lemmatize the input word.

3.1.2. GIS pipeline

This early version of Finger simply queries GeoNames' public API with the lemmatized wordform, and if the query is successful, inserts the best answer's coordinate pair to the output dataframe. No further disambiguation is done on Finger's side. If the query is unsuccessful, a Python None object is inserted instead. A Python module named GeoCoder[9] is used to query GeoNames and handle the returned object. The advantage of this method is that the user does not have to download any gazetteer

[7] https://voikko.puimula.org/
[8] https://joukahainen.puimula.org/
[9] https://geocoder.readthedocs.io/

data and the database is always up to date. The web service also does some sort of ranking, meaning that it returns Helsinki, the capital, before Helsinki, the village. This is evidenced by the API's relatively strong performance in Karimzadeh et al.'s (2019, pp. 12–13) results. The exact parameters of this ranking process are not, to my knowledge, published. GeoNames' online API requires each user to create a free account, which is used as an API key, and it is rate-limited to about 1000 queries per hour[10].

Nonetheless, I chose GeoNames and the API approach over others for a few reasons. First, the online API makes setting up the geoparser easier, since the user does not need to download database files locally. Second, GeoNames has global spatial coverage, unlike NLS' Finnish place names, and primarily includes Finnish place name variants (like *Lontoo* when referring *London, UK*), unlike TGN. Because I presumed toponym mentions in Finnish text would be weighed towards Finland, I briefly explored the spatial coverage completeness of GeoNames against NLS place names in Figure 10. Coverage comparison of GeoNames features (database dump from May 2021) and National Land Survey placenames (1:20 000 scale, year 2020) for Finland and Åland. as inspired by a global comparison in Acheson et al. (2017). While the national dataset is of course more complete in terms of spatial and thematic coverage, I find GeoNames sufficient for this purpose. The toponyms are naturally not completely uniformly spread spatially: notice, however, the curious cluster in Southern Finland near Kouvola on GeoNames' side. On that particular area, a much greater portion of toponyms, including those of local geographical features are included – the reason is unknown to me, but since GeoNames allows for user input, it might be the work of a dedicated local. This highlights one of the weaknesses of a (partly) crowdsourced gazetteer.

[10] https://www.geonames.org/export/

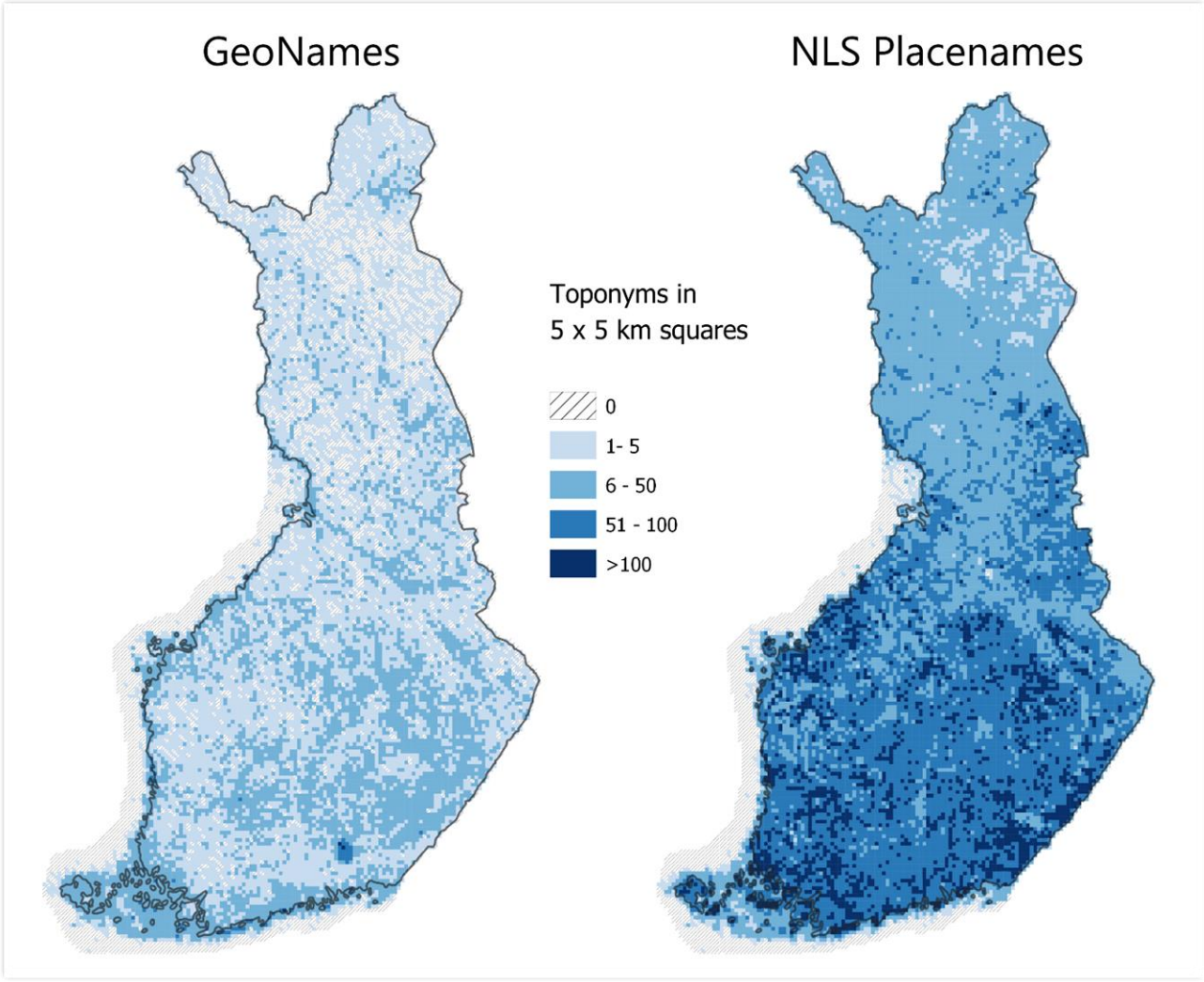


Figure 10. Coverage comparison of GeoNames features (database dump from May 2021) and National Land Survey placenames (1:20 000 scale, year 2020) for Finland and Åland.

3.1.3. Finger output

After running the two pipelines, Finger outputs the final, formatted collection of possible toponyms and point coordinates. The user is given two options: a JSON array or a Pandas (Reback et al., 2021) DataFrame. Pandas is a widely used Python library for data processing that Finger also uses internally. DataFrame is the standard selection that allows the most flexibility in terms of output length and options: it is represented in Figure 9. The exact features it output in the dataframe are described in Finger's repository[11]. In addition, Finger allows for outputting JSON arrays as defined in J. Wang and Hu (2019b) and implemented in the EUPEG platform (for an example, see J. Wang & Hu, 2019b, p. 20). This option is included for compatibility with the evaluation platform and previous geoparsers.

[11] https://github.com/Tadusko/fi-geoparser

3.2. Creating evaluation data

Finger had to be compared against human-verified data to evaluate its performance. To my knowledge, no publicly available Finnish geoparsing corpora exist. Therefore, suitable texts had to be obtained and the toponyms *annotated* with tags and coordinates. Two new test datasets were created: Finger-news and Finger-tweets. The former consists of Finnish Wikinews (Wikiuutiset)[12] articles and the latter of tweets where the primary language is Finnish. I selected these sources because they represent different text genres and because they were used in similar English corpora: GeoVirus (Gritta et al., 2018) and GeoCorpora (Wallgrün et al., 2018), respectively. All the code used in the following tasks and the final corpora are shared in a GitHub repository.[13]

First, I acquired the texts from online sources. I collected a list of all Wikiuutiset article URLs from the year 2011 (n=42) and scraped the article texts and titles with a Python script. The input texts consist of the titles followed by the article texts. For Finger-tweets, I queried Twitter API in late August 2021 for Finnish tweets using twarc Python package (Summers et al., 2021). Because the query could not be empty, I used a neutral conjunction *ja* (and) as the search string. One thousand tweets were randomly sampled from the tens of thousands returned by the query: temporally, all the tweets are posted within a few days in August 2021.

Next, the input texts were annotated by me and two other students. A geoannotation tool where the annotator could mark the correct locations on a map window within the program would prove useful in this step. Although such a tool was developed by e.g. Karimzadeh and MacEachren (2019), installing and running their tool turned out to be challenging. Instead, I used a Python-based general annotation tool Label Studio (Tkachenko et al., 2020). Each toponym in each document was labelled with a LOC tag, similarly to the example in Figure 4. A document may have zero toponyms, it is nonetheless still included in the corpus. For location information, the annotators primarily queried GeoNames' web portal[14] and copied the latitude-longitude coordinate pair to each toponym's metadata. If the toponym was missing in GeoNames, the annotators queried alternative gazetteers: NLS Geographic Names through Nimisampo interface[15], OpenStreetMap[16] or Google Maps[17] in this

[12] https://fi.wikinews.org/
[13] https://github.com/Tadusko/finger-corpora
[14] https://www.geonames.org/
[15] https://nimisampo.fi/en/
[16] https://www.openstreetmap.org/
[17] https://www.google.com/maps/

order. In case the coordinates were nonetheless unavailable, e.g. because the toponym refers to an inexact or outdated location, a NaN tag was used. There are 4 such cases in Finger-news and 14 in Finger-tweets.

Different toponym definitions were used when annotating the two corpora, mimicking the definitions used in GeoVirus and GeoCorpora. As listed by J. Wang and Hu (2019b, p. 11), only administrative units (such as countries and cities) are marked in GeoVirus – this schema is used in Finger-news. Additional features such as buildings and natural features are annotated in GeoCorpora and Finger-tweets. For example, *Kuopio* is annotated in both, but the lake *Saimaa* only in Finger-tweets. Different annotation schemas were selected to acknowledge the plurality of toponym definitions in the preceding research and to explore how they affect the geoparser's performance.

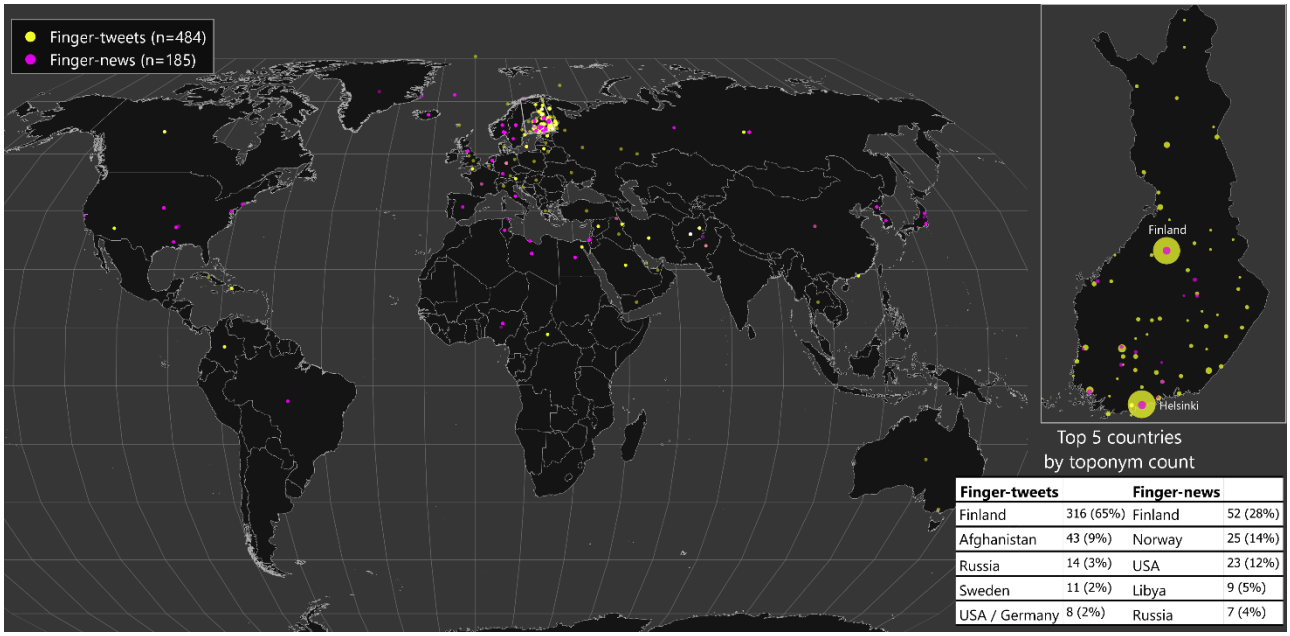
There were multiple other considerations on what to annotate, such as the difficulty in embedded toponyms: e.g., should France be annotated in the French Revolution. More details on the annotation decisions made in this work are listed in Appendix A. While annotating the tweets, 20 tweets were found to not be in Finnish – they were dismissed from Finger-tweets, leaving the final document count to 980. See this and other numeric descriptions in Table 4. Note, for example, that a single news article predictably has, on average, a lot more toponyms than a single tweet. In fact, only 285 tweets of the total 980 have any toponyms.

Table 4. The Finger corpora in numbers. Total tokens tells how many words are in the dataset: a token is a span of text separated by whitespace (for example, New Delhi is two tokens although it is a single toponym)

Dataset	**Documents**	**Total tokens**	**Total toponyms**	**Mean toponyms per document**
Finger-news	42	6352	189	4.5
Finger-tweets	980	22,513	498	0.51

Only coordinates were marked in the corpora instead of persistent identifiers, such as feature ID codes: this was done to lessen workload and because the coordinates came from many different gazetteers. That is why no comprehensive listing by location type (nation, city, natural feature etc.) is easily available. Examining the toponyms manually shows they skew heavily towards political, such as administrative areas, even in Finger-tweets where more entity types were annotated. The spatial spread of the toponyms is examined in Figure 11, which lists the most frequent countries the

toponyms fall in. Finland is the most frequent one in both. However, there are differences: in Finger-tweets the skew towards Finland is much more pronounced. The tweet texts also contain a high number of references to Afghanistan, which explained by the Taliban takeover in August 2021 occurring simultaneously to the tweet collection.



Finger-tweets		Finger-news	
Finland	316 (65%)	Finland	52 (28%)
Afghanistan	43 (9%)	Norway	25 (14%)
Russia	14 (3%)	USA	23 (12%)
Sweden	11 (2%)	Libya	9 (5%)
USA / Germany	8 (2%)	Russia	7 (4%)

Figure 11. Spatial distribution of the known locations in Finger-tweets and Finger-news globally and in Finland: locations of toponyms and top 5 countries by count in the table.

The corpora and the related code used to acquire and format the corpora are shared in a GitHub repository[18]. The licensing rules of Wikinews and Twitter add some restrictions on the level of openness the resources can be shared in. Finger-news is shared as-is, but the public version of Finger-tweets is stripped of the tweet texts, which are replaced with the tweet identifiers. The IDs may be used to query the original tweets, provided the tweets have not been deleted – deletion is a problem with such social media corpora, as noted by Gritta et al. (2020, p. 699). Some other caveats should be considered when using the corpora. First, the corpora are not divided into dev-train-test sets due to their small size: they are primarily meant for testing, not training Finnish geoparsers. Second, no inter-annotator agreement was measured. Usually when building a corpus, a bit of the input is annotated by multiple annotators. The amount of overlap in the annotations, or agreement, between annotators is calculated – low agreement can indicate that the task is poorly defined, or the annotation

[18] https://github.com/Tadusko/finger-corpora

schema flawed. However, overlapping annotation was not done in this work due to the scope and focus of this thesis.

The corpora are formatted in a XML file according to the standards set in the EUPEG platform (J. Wang & Hu, 2019b). XML allows for a nested structure of multiple *entries*, or documents, followed by zero or more toponyms. Each toponym, in turn, has several attributes, such as the toponym start and end indices in the input text, and WGS84 coordinates in longitude-latitude order. This is the minimal information required for a geoparsing corpus and the one used in these corpora, but the format allows expanding the entries with any necessary information, such as alternative names or identifiers. See J. Wang and Hu (2019b, pp. 18–19) for full explanation and Listing 1 for an example from this work.

```
<?xml version='1.0' encoding='utf-8'?>
<entries>
    <entry>
        <text>Savonlinna ja Mikkeli on tosi kivoja!</text>
        <toponyms>
            <toponym>
                <start>0</start>
                <end>10</end>
                <phrase>Savonlinna</phrase>
                <place>
                    <footprint>29.10818 61.8624</footprint>
                </place>
            </toponym>
            <toponym>
                <start>14</start>
                <end>21</end>
                <phrase>Mikkeli</phrase>
                <place>
                    <footprint>27.33025 61.64117</footprint>
                </place>
            </toponym>
       </toponyms>
    </entry>
    . . . . .
</entries>
```

Listing 1. The XML format Finger-news and Finger-tweets are shared in.

3.3. Additional evaluation metrics: lemmatization and query errors

In addition to the evaluation metrics discussed in Section 2.6, I believe two more measures are necessary to gauge Finger's performance accurately: lemmatization and query errors. Both of these are linked to lemmatization, which is a more prominent issue in Finnish than in English due to the former's more complex morphology, where the toponyms can be conjugated in many ways. Almost every toponym Finger encounters must therefore be lemmatized. Lemmatizers can provide erroneous

lemmas, such as *Alankomaissa* → *Alankomaa* instead of *Alankomaat*. A lemma error flows downstream since a gazetteer query with an incorrect lemma will likely either fail or resolve incorrectly.

The first metric, which I call *lemmatization error*, is simply the ratio between erroneous lemmas and the true positives. If the count of erroneous lemmas is EL and count of toponyms correctly recognized by the system (true positives) is TP, then:

$$\text{Lemmatization error} = \frac{EL}{TP}$$

A lower value indicates better performance, that is, fewer erroneous lemmas. I count the erroneous lemmas by hand, since there are no gold-standard lemmas that would enable automatic checking. I do not count cases where the toponym has a typo (like *Afganisthan* instead of Afghanistan or Afganistan) as lemmatizer errors, only cases where the lemmatizer could be expected to function correctly.

Not all correctly recognized toponyms get resolved to locations. In the previous research, this is mainly caused by toponyms missing from primary gazetteers, causing the query to return zero candidates. This can be dealt with in different ways: for example, Gritta et al. (2020) did not evaluate the toponyms which were missing from GeoNames and had to be annotated with Google Maps. They do, however, report the number of toponyms resolved (Gritta et al., 2020, p. 703). Similarly, I report the share of successful toponym queries, which also reveals how many failed. I do this because lemmatization and missing entries in GeoNames could lead to significant omissions in the final results no matter how well the toponyms were recognized.

If *Resolved* is the count of toponym queries that were successful (that is, not null) and TP is the count of true positives, then:

$$\%\ Resolved = \frac{Resolved}{TP}$$

4. Results

4.1. Recognition and resolution evaluation on the Finger corpora

Finger's toponym recognition and lemma error results are reported in Table 5. As the F-score shows, the geoparser performs similarly on both datasets, though slightly better on the tweet dataset. The

geoparser recalls roughly 4/5 of all toponyms, but its precision is worse on both corpora. There is a marked difference in lemmatizer performance between the corpora: about 1/5 of the correctly identified toponyms (true positives) in Finger-news were incorrectly lemmatized, while both absolutely and relatively this number is much smaller for Finger-tweets.

Table 5. Toponym recognition evaluation results. Evaluation measures for an exact match: partial matches are counted as errors. The arrows indicate whether lower ↓ or higher ↑ value is better.

Dataset	**Precision ↑**	**Recall ↑**	**F-score ↑**	**Lemmatization errors ↓**
Finger-news	0.701	0.794	0.745	$^{32}/_{150}$ (21.3 %)
Finger-tweets	0.717	0.819	0.765	$^{28}/_{408}$ (6.9%)

Toponym resolution evaluation is presented in Table 6. Over half of the input toponyms were correctly located on both cases, as indicated by the median error. A vast majority of the toponyms were resolved within 161 kilometers on both corpora. Metrics that highlight significantly large errors – mean error and Accuracy@161 km – are both worse on Finger-news. Conversely, area under the curve is about equal for the two. Because the error distances were transformed by taking the natural logarithm of them, smaller error distances get emphasized in AUC, which indicates Finger-tweets had a larger share of small errors. Altogether, the results indicate that Finger generated larger resolution errors on the news corpus and perhaps more, but smaller on the tweet corpus. This interpretation is supported by Figure 12, where the error curves are plotted. Notice that the curve for Finger-news is steeper than the more gradual rise present for Finger-tweets. Finally, a significant portion of the input toponyms did not get resolved at all, most likely due to a lemmatization error: this share is larger for Finger-news, as expected due to the lemmatization errors (Table 5).

Table 6. Toponym resolution evaluation. These results are for the correctly recognized toponyms (true positives) that contain coordinates and which were successfully resolved (i.e. the query did not return an empty result). The arrows indicate whether lower ↓ or higher ↑ value is better.

Dataset	Mean error (km) ↓	Median error (km) ↓	ACC@ 161 ↑	AUC ↓	% Resolved ↑
Finger-news	286.1	0	0.917	0.083	$^{123}/_{150}$ (82.0 %)
Finger-tweets	127.2	0	0.958	0.093	$^{369}/_{408}$ (90.44 %)

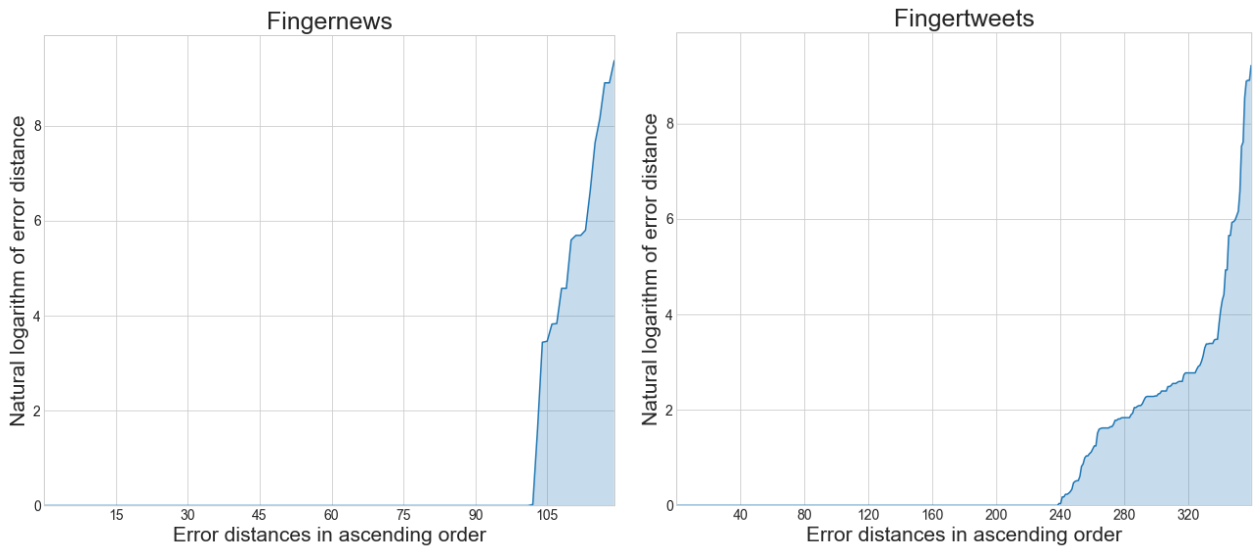


Figure 12. AUC error curves. Finger-news left, Finger-tweets right. Note that the Y axis is on a logarithmic scale.

4.2. Error analysis

In this section, I show examples of the kinds of errors Finger makes. These can be divided to the toponym recognition errors (false negatives and false positives), lemmatizer errors and toponym resolution errors. For the recognition errors, it is somewhat hard to know exactly what features the classifier uses to classify as a location; the deep neural models are oblique in that way. However, some patterns can be deduced from the output.

Some of the false positives are explained by the classifier falling for geo/non-geo ambiguity: it marked proper nouns like Yhtenäinen Venäjä (*United Russia*, a political party) and Taliban, and common words such as UTC-aikaa (*UTC time*) and kanssa (*with*). Hashtags in the tweets proved

tough for the classifier: they are frequently inserted at the end of tweets to indicate the spatial area of the tweets and often with non-standard spelling. The classifier seemingly uses capitalization as a feature to recognize toponyms, but it is not an absolute requirement – some toponyms written all lower case and even ones containing typos are recognized. Adding to false positives, the classifier is also sensitive to emojis 🙄 and styled Unicode text 𝓵𝓲𝓴𝓮 𝓽𝓱𝓲𝓼. While emojis get filtered by Finger because of their short length, the examples indicate that the classifier might be overly greedy on random noise – perhaps because these characters are rarely present in the NER layer training corpora or the BERT pre-training material. Boundary errors are frequent: for example, Irakin Kurdistan (*Iraqi Kurdistan*) and Härmän käräjätalo (*Härmä courthouse*) are annotated as one toponym, but the NER classifier marks the toponyms separately in the first example and marks Härmä in the second. These results would not be useless in a real use case, but they are more imprecise than the correct interpretation. Boundary errors are also partly a matter of interpretation

Because it does a simple dictionary lookup, the lemmatizer is easy to predict: if it makes an error once, it will make the same error every time. This leads to some systemic errors, such as incorrect lemma for Yhdysvaltain → Yhdysvalta (**the United State*). The reason may even be looked up at the online version of Joukahainen dictionary[19]: in this case, *Yhdysvallat* has no conjugation paradigms set, which is why the lemmatizer falls back on the assumption that the plural form should be returned to singular. These systemic errors linked especially to a few country names probably explain why the share of lemma errors is so significant in Finger-news (Table 5). Some toponym phrases are lemmatized incorrectly because the words are treated individually instead of as a phrase: for example, Englannin kanaalissa → Englanti kanaali (**the England Channel*). Finally, the lemmatizer was unable to deal with cases with non-standard spelling: e.g., hashtags where a postfix is added after a hyphen, such as #Kouvola'ssa.

Lastly, there are toponym resolution errors. A toponym might either be incorrectly located or completely unresolved (the gazetteer query fails to return a single candidate). Not all toponyms located differently from the gold locations are actual errors, they might be differently interpreted. For example, GeoNames frequently includes Finnish municipalities twice under the same name: once as a point representing the whole municipality and once as the "seat of the administrative division" (*alueen keskus* or *taajama*). See for example the query for Pälkäne[20]: two points some kilometers away, but both would be usable in an actual use case. Then there are actual errors, such as falsely

[19] https://joukahainen.puimula.org/word/edit?wid=522364
[20] https://www.geonames.org/search.html?q=p%C3%A4lk%C3%A4ne&country=

resolved geo/geo ambiguity when *Vuores* is grounded to Sweden instead of the town in Finland. The query might fail completely: the reason is either an erroneous lemma or the toponym missing from GeoNames. Omissions include Finnish facilities (Musiikkitalo, Hahkialan kartano) and addresses (Myllymatkantie), which seem to be out of the gazetteer's scope.

Drawing the errors together, the linguistic and geographic ambiguities discussed in Section 2.3 certainly present when geoparsing Finger-tweets and Finger-news. In addition, there are errors caused by the poor performance of the lemmatizer and by the classifier's inability to deal with unexpected input types, like emojis. Some of the differences, like boundary errors and variants of nearly the same locations, might not be major errors in actual use cases.

5. Discussion

At the beginning of this thesis, I wondered how reliable and usable is Finger? By reliability, I refer to the confidence that can be placed on the geoparser fulfill its purpose: to find place names among texts and correctly locate them. This is what most of the thesis has aimed to test. Usability is a tougher topic and one that I can only discuss in this thesis: even if geoparsing is successful, what can the results be used for – can they provide answers to relevant questions? What is the level of development of the current system and what I believe should be the next improvements for Finger and geoparsing at large: these are the topics addressed in this chapter.

5.1. Contextualizing the results

The results show that Finger performs about equally well on both corpora, even slightly better on Finger-tweets (Table 5 and Table 6). This is puzzling, because my initial hypothesis was that news and tweets were from sufficiently different domains to result in differences in performance – the short tweets, which include colloquialisms, being the harder dataset (see Section 2.6.3). I also presumed tweets might include more references to local events and thus more fine-grained, local toponyms. As well, the NER corpus used in training lacks tweets and similar short-form social media content (Luoma et al., 2020). Moreover, I used a different annotation schema: I presumed annotating more and rarer toponyms (referring to roads, buildings etc.) would affect recall on Finger-tweets. This would have mirrored the performance differences on English news and tweet geoparsing corpora (J. Wang & Hu, 2019a).

Several reasons for this difference can be proposed. First, language use in Twitter is not as different from news as I presumed: during annotation, we noticed that a clear minority of the toponyms contained colloquialisms or varying capitalization. Shared news excerpts, posts by organizational accounts and such bring tweets' language form closer to that of news articles. Second, it could be that the narrower annotation schema worked against Finger in this case. As mentioned in Section 3.1.1, the Annotation guidelines for the Turku NER corpus[21] show that buildings, landmarks, and natural features are included under location, making the schema more encompassing than the one in Finger-news. The mismatching schema led to toponyms such as *Pohjoinen jäämeri* (the Arctic Ocean) being marked as false positives, which somewhat misleadingly lowers precision. This showcases the importance of clear toponym definitions: you find what you annotate. Finally, the test sets are small, consisting only of 42 news articles and less than thousand tweets, of which only some contained toponyms (Table 4). Smallness makes them more susceptible to random variation: that is, a phenomenon that would have surfaced in a larger study did not in this case.

While the reasons listed above could explain the roughly equal performance on the two corpora, the outcome is of course not a bad one. An alternative interpretation is that Finger can process these two text types at a sufficient level. But are the results robust overall? The results reported in Chapter 4 do not, in isolation, directly answer the question of whether they are good or not: it is not completely clear what even is good performance. Because no equivalent research on Finnish datasets has, to my knowledge, been done previously, the results are tough to contextualize. To do so, I will draw from two English geoparser performance reviews (Gritta et al., 2017; J. Wang & Hu, 2019a).

J. Wang and Hu (2019a) evaluate eleven geoparsers on numerous geoparsing corpora for English on the same metrics as used in this thesis. The geoparser implementations range from rule-based, like the Edinburgh geoparser (Tobin et al., 2010), to the latest neural geoparsers (X. Wang et al., 2019). Among the corpora evaluated are GeoVirus and GeoCorpora, which are the model corpora for Finger-news and Finger-tweets respectively, though naturally not strictly equivalent. For example, GeoVirus focuses on disease news (Gritta et al., 2018) while Finger-news has no thematic branding; GeoCorpora used various keyword filters and spread out the tweets temporally (Wallgrün et al., 2018), while Finger-tweets were not thematically filtered and are all within a few days timeframe.

Included in the geoparsers is one based on an older version of spaCy's NER engine for English and population heuristic for toponym resolution. It does not use a neural NER classifier, and the population baseline is not exactly the same as the GeoNames web service query used here. Despite

[21] https://github.com/TurkuNLP/turku-ner-corpus/blob/master/docs/Turku-NER-guidelines-v1.pdf

the clear differences, I believe it to be the closest simile to Finger, which is why it is the starting point for contextualizing the results. Some key figures to support the following discussion are presented in Table 7. spaCy + population performed among the worst in toponym recognition in both GeoVirus and GeoCorpora (J. Wang & Hu, 2019a, pp. 4–5); the difference is more pronounced in the generally easier news corpus GeoVirus. If we assume the English corpora can be compared to their Finnish counterparts, Finger's recognition performance would be placed somewhere in the lower middle for GeoVirus and at the top for GeoCorpora. Overall, the tweet corpus was a bigger challenge than the news corpus for the geoparsers; a result that did not occur in this thesis.

Table 7. Geoparser performance excerpts from J. Wang and Hu (2019) on the left and this work repeated on the right. F-score and Area Under the Curve (AUC) represent toponym recognition and resolution results, respectively. The arrows indicate whether lower ↓ or higher ↑ value is better.

	Results for spaCy and best performing geoparser (in parentheses)	
Corpus	F-score ↑	AUC ↓
GeoVirus	0.499 (0.917)	0.367 (0.319)
GeoCorpora	0.562 (0.763)	0.224 (0.084)

Results for Finger		
Corpus	F-score ↑	AUC ↓
Finger-news	0.745	0.083
Finger-tweets	0.765	0.093

Toponym resolution results are significantly better for Finger than almost any of the results reported by J. Wang and Hu (2019a): see for example the difference in AUC scores in Table 7. It is, however, crucial to remember that resolution is usually measured only for the toponyms recognized in the previous step. This means that potentially challenging toponyms, and their large error distances, get filtered out by before evaluating them, especially when recall is low. This lowers the averaged error distances: this effect was, for example, demonstrated by Karimzadeh et al. (2019, p. 12), who evaluated toponym resolution with and without the preceding step. I also hypothesize that there is more potential for large error distances in the global English corpora – this is due to the size of the English-speaking world. Similarly to how a majority of the toponyms in the Finger corpora fall within Finland (Figure 11), the significant portion of toponyms in e.g. GeoCorpora fall within United States and the rest of the Anglosphere (Wallgrün et al., 2018). Since the metrics are sensitive to distance, I believe the task might be easier on the Finnish corpora, owing to the sheer difference in geographical

area, which also increases the number of toponyms and potential geo/geo ambiguity. This is, however, only speculation on my part.

In total, then, Finger performs robustly side-by-side with the English geoparsers; though I must emphasize that the comparison cannot be a direct one, since the results are acquired from different geoparsers run on different corpora. Beyond the numbers, there remains a question of whether the current geoparsers are valid as sources of location information. Both Gritta et al. (2017) and Wang and Hu (2019a) have addressed this question. Gritta et al. (2017) concluded that the current English-language geoparsers they reviewed are ill-suited for sources of data, due to their limitations and the complexity of the task. They argue that geoparsing could be used as an auxiliary data-source, whose outputs cannot be used without a critical eye on their limitations. J. Wang and Hu's (2019a) more recent assessment concludes that it depends on what the exact task is: on use cases such as international news geoparsing, the task may very well be considered solved due to the to the overall high performance of the best systems. However, using geoparsers on the more challenging datasets (see Section 2.6.3 for discussion), many of the simpler geoparsers stop working. Thus, there is still need for user oversight on designing the task and supervising the geoparser's output.

The same cautions apply to Finger as well. Consider that Finger recalled roughly 80 % percent of toponyms in the corpora and of those, 82–90 % got resolved at all (Table 5 and Table 6). In concrete numbers, that means Finger returned $\frac{123}{189}$ locations for Finger-news and for $\frac{369}{498}$ the tweets. That is the amount of data Finger misses. The system also adds noise in the data, as falsely identified toponyms are added in the output (64 for Finger-news and 161 for Finger-tweets). Both of these should be considered when using the system. Toponym resolution is another possible source of error. Finger is still a long way from a level of performance that a human would perform in. The results are further influenced by lemmatization, which introduces a source of error not present in the evaluations of English geoparsers. At its current state, I believe Finger can be used as a first step in the geographical analysis of texts, but its outputs should be critically examined by the user before applying them to any downstream tasks. However, like the previous assessments (Gritta et al., 2017; J. Wang & Hu, 2019a), I am hopeful that geoparsers can be further improved.

5.2. Developing Finger: more, and better

Finger is a developing program and its current status only a starting point. Finger and Finnish geoparser research can be advanced on multiple fronts: the lemmatizer errors can be mitigated with a

different lemmatization method, the toponym recognition and resolution solutions can be developed and new corpora created to allow for training and evaluation, as I elaborate below.

As discussed in the error analysis (Section 4.2), lemmatization errors are quite frequent, which affects the succeeding task of resolving the toponym. The quality of lemmatization is therefore crucial for Finnish and similar morphologically rich languages, insofar they rely on gazetteers which contain toponyms in their base forms. While no solution is perfect, many of Finger's lemmatization errors can be traced to the current look-up based lemmatizer, which I presume might be ill-suited for lemmatizing toponyms, which take varying surface forms. Having a fixed dictionary is limiting especially when the input is free-form internet texts. In addition, it would be beneficial to have a context-aware lemmatizer instead of lemmatizing every token individually like currently: such as system could better deal with e.g. phrases like *Englannin kanaali* (Kanerva et al., 2021, pp. 545–546). Kanerva et al. (2021) implemented a neural lemmatizer trained on Universal Dependencies that performed better than baseline systems. A similar implementation could be explored in Finger and implemented in the spaCy NLP pipeline. An additional benefit would be an easier installation, since currently the lemmatizer requires users to install an additional Python library and download the related dictionary.

I propose that the general-purpose neural NER tagger using contextual word embeddings is a robust solution, and one that can be incremented on instead of replaced. Firstly, the tagger could be re-trained with a more recent NER corpus that combines the two corpora discussed previously (Luoma et al., 2021). This combined corpus is also mapped to a different annotation schema, OntoNotes: this schema contains facilities (e.g. buildings), geopolitical entities and natural places annotated with separate tags. This information is useful in geoparsing: the user could, for example, decide they are only interested in natural locations, such as mountains, and the rest could be excluded. Secondly, multilingual language models could be explored. Neural language models, like BERT, can be trained materials from one language, such as the monolingual Finnish BERT used in this thesis, or multiple languages. One such multilingual model is the Finnish-English BERT model, which almost reached the performance levels of monolingual models when evaluated (Chang, 2021). I see potential in such models because they could allow for more language independency in the inputs – for example, just a bilingual model for Finnish and English could handle the vast majority of tweets posted in Finland (Hiippala et al., 2020). Third, recent geoparsing research has proposed architectures suited specifically for toponym recognition; I believe Finger could benefit from these. For example NeuroTPR, which embeds not only words but also characters in an attempt to better recognize toponyms in social media posts (J. Wang et al., 2020).

A simple gazetteer query is a decent baseline system for toponym resolution: it is easy to implement and use, and mostly works for the obvious toponyms. Such a system will, however, never reach human-like performance. A single query will always resolve to one location – always the capital and never the town – even when the context would provide ample evidence otherwise. I therefore believe it is imperative to implement a more sophisticated resolving pipeline in Finger. Mirroring the developments in toponym recognition, the latest disambiguation solutions are built on neural models that model the toponyms beyond the surface forms. These exploit lexical features, like toponyms and context windows (Cardoso et al., 2019; Kulkarni et al., 2020) and/or geographical features (Gritta et al., 2018) modeled as embedding vectors. These solutions may also be used independent of gazetteers (Kulkarni et al., 2020). Because of the strong reported results, a future toponym resolver in Finger should be based on similar techniques.

Lastly, bigger and better geoparsing corpora are needed to accurately measure the performance of and allow for further development of Finnish geoparsers – a corpus can be used for testing and training. The corpora discussed in this thesis employed one of three approaches: manual annotation by a few experts, often the authors, crowdsourcing most of the work through marketplaces such as Amazon's Mechanical Turk or automating the process. Manual annotation mostly costs time, but large corpora require a prohibitive amount of work for a small team of annotators: this tactic was used in this thesis and e.g. the Local-Global Lexicon (Lieberman et al., 2010). GeoCorpora was first annotated by paid workers on a crowdsourcing platform and refined by experts later on, which allowed for the annotation of over 6000 tweets (Wallgrün et al., 2018). Lastly, multiple articles propose processes that automatically acquire sentences with toponyms and coordinates: these are primarily based on Wikipedia, which (at least the English Wikipedia) contains hyperlinks and coordinates for articles about places (Laparra & Bethard, 2020; J. Wang et al., 2020).

All of these approaches could be explored creating further geoparsing corpora for Finnish. Automated methods hold much promise, because using them could circumvent the costly (either time or funds) annotation process and allow for creation of large datasets. A sufficiently large, high-quality corpus would be ideal for training the next-generation of Finnish geoparsers. On the other hand, manual annotation would allow for a more nuanced toponym annotation schemes, such as exploring the literal vs. associative toponym taxonomy (Gritta et al., 2020) or nested named entity tagging, where named entities are allowed to contain named entities (*University of Turku* [ORG] contains *Turku* [LOC]) (Ringland, 2016). Whichever approach is used, the definition of a toponym, must be clearly stated in the annotation schema. This is because the toponym definition flows downstream from the corpus to the NER tagger, which in turn affects any geoparser trained or tested on the corpus.

5.3. Beyond it all: next steps for geoparsing

Geoparsing is an active field of study. In the previous sections, I presented some suggestions on how to evolve Finger on a technical level. Those advancements would apply to geoparsing as it is described in this thesis – however, I believe geoparsing as a task will mature. The task has potential to go beyond the concepts presented here: beyond one language, beyond just toponyms, and representations of space beyond coordinate points, as I elaborate below.

This thesis has focused on Finnish geoparsing and discussed the issue mostly through research done on English: the language choices were directed by my language capabilities, of course, but also on what the previous research has focused on. That is, English as a language of study. The problem of lemmatization has been discussed at length, but other differences between the *lingua franca* and smaller ones are worth pondering. For example, gazetteers that, often by default, include the English names for places must also have variants for the target language to be usable. What about how English may divert the geographical focus to the Anglophone world or how much easier research might be using large and richly annotated English Wikipedia in comparison to smaller languages? Recognizing these, I believe geoparsing research ought to adopt the Bender rule from computational linguistics. It goes like this: "Always name the language(s) you are working on" (Bender, 2019). The central point of the rule is to recognize that English is a language among others, and it is possible that techniques developed for it are language-dependent or divert research someplace that is not widely applicable. For geoparsing, then, when asking questions like *is geoparsing a solved problem*, I believe it should be appended with *is geoparsing a solved problem for* ***English / this language*****?**

This thesis focuses on toponyms as linguistic descriptors of space. The implicit assumption here is that, in a sentence with a location reference, the toponym reveals that location. But then think of a phrase like *100 km west of Helsinki*, where the toponym is merely a *landmark* used to structure space. These types of phrases can describe many forms of relationships, such as adjacency, containment and distance between objects (Stock & Yousaf, 2018). To resolve these location referents, recognizing the phrases and understanding how to transform them is crucial: if *100 km west of Helsinki* is recognized, where should the resolved location be placed? To this end, corpora containing these phrases have been created (Laparra & Bethard, 2020; Stock et al., 2021). In geoparsing, I believe the important question is how much would processing these affect the location information vs. a baseline system of geoparsing as usual? In which use cases would such a system bring additional benefits?

If toponyms are a simplistic way of viewing how space is described through language, so is a coordinate point an abstraction of geographical space. How crude of an abstraction depends on the

scale that is relevant for the use case (more on this later). While the coordinate points can be grouped and clustered in different ways (Hu, 2018b), nonetheless, it is evident that, for example, a continent represented by a point tells very little of the extents of that location. A point, as a zero-dimensional abstraction, also does not tell anything about the relative size differences between, e.g., an airport and a nation: such differences would have to be communicated through attribute data. Laparra and Bethard's (2020) corpus GeoCoDe contains polygons and lines alongside points. They also present a metric that measures the amount of areal overlap, instead of the error distance-based metrics presented in Section 2.6.2. While a polygon is in many cases more accurate representations of locations than points, they are also more demanding computationally and still have varying levels of precision. For example, the National Land Survey of Finland offers the municipal borders of Finland dataset[22] in five different scales, or spatial resolution, each one generalized more than the last. My point is that polygons are still abstractions with their strengths and weaknesses.

An interesting direction is to partition the globe into hierarchical cells (Adams, 2017). The grids contain cells in different levels, from exact (e.g. 1 km^2) to crude, and the fine-grained cells are contained within the general ones. Several deep learning toponym resolvers use such grids to predict the likeliest location, but then transform that prediction to a coordinate point (Cardoso et al., 2019; Gritta et al., 2018; Kulkarni et al., 2020). I wonder if these grid cells could be returned to the user as-is; smaller cells to represent fine-grained locations and very general ones for, e.g., nations. The grid cells would not be significantly more computationally complex than points, they would approximate the location's scale and fine-grained cells could be aggregated to larger ones should the analysis require it. Continuing on this line of thought, I wonder if a type of probability surface could be used to inform the user of locational impreciseness when the location lacks clear boundaries (vague cognitive regions, like *Stadi*).

Returning to the topic of scale and granularity (discussed in Section 2.3.2), I believe it should be addressed in future geoparsing research. This is for a few reasons. First, the point-based toponym resolution error metrics are not sensitive to the scale of the location resolved. For example, let us say a city is resolved to a point that is 20 km away from the gold location, perhaps due to different gazetteers or multiple similar entries in one gazetteer. This error is not significant for a human interpreter. But what if the toponym geoparsed is a fine-grained one, like a university campus in the city and it is resolved to a restaurant in the other side of the city 20 km away? Current metrics treat these cases the same, although the latter is a serious misplacement and the former simply a different

[22] https://www.maanmittauslaitos.fi/en/maps-and-spatial-data/expert-users/product-descriptions/municipal-division

point abstraction of a large area. I thus agree with Gritta et al. (2017, p. 618), who call for an error metric that scales in relation to the size of the location. The functional operating scale of geoparsers is something that the users should be aware of as well: while some research has explored sub-city scale geoparsing (Alex et al., 2019; Rocco et al., 2021), most of the research presented in this thesis operates at granularity level of cities and nations. Finger is no different: while there is no theoretical reason it could not be applied to, for example, the exploration of sports facilities at a municipal level akin to Koivisto (2021), it is limited by gazetteer choice, and an evaluation metric like Accuracy at 161 km is not primed to find fine-grained errors. I believe, then, that geoparsers should have an explicit operating scale, or operating modes, and that those are clearly communicated to the user.

While I have discussed different nuances of the problem of geoparsing at length, the final goal of geoparsing should not be lost – geoparsing is a crucial step in research processes that apply geographical analysis to texts: consisting of data retrieval, geoparsing, analysis and visualization (Hu, 2018b, p. 11). Thus, geoparsing is given relevance and usefulness through applied research. To wrap up this thesis, I would like to propose some potential use cases. Geoparsing could be explored as a method to geolocate social media users (Zheng et al., 2018) based on their location mention history, in the same vein as geotagged post history: and measure for example, if an accurate population distribution pattern would emerge similarly to users located with geotags (Järv, 2020). The potential here would be to geolocate more users and use that information in whatever succeeding research. The conceptual separation between being somewhere and talking about somewhere (from/about data, see (Hu, 2018b)) could be explored through a massive geotagged social media data repository, such as Hiippala et al. (2020). When a geotagged tweet contains toponyms, to what extent are the geotagged locations related to the geoparsed locations? Could from/about data be separated from one another and input texts classified in either class? To give a specific example, it would be interesting to explore whether visits to national parks (Heikinheimo et al., 2017) could be quantified through geoparsed social media content. Or, whether there are differences in the semantics of the content when a user visits a place versus discusses it online. These topics and more could be addressed in future research: lessons learned from the applied use cases flow back to geoparser research and Finger's development.

6. Conclusion

This thesis aimed at adding to the methodical toolbox of geoinformatics by exploring the topic of geoparsing: recognizing toponyms in free-from texts and acquiring the correct spatial representation for them. A geoparser could, when functioning correctly, offer means of acquiring new, semantically

rich geodata: if the locational component is reliable, the attribute data contained in, for example, social media posts offers many opportunities. Such a system shows promise for a wide assortment of fields from geographers looking for a new way to understand human interaction and quickly emerging events to those attempting to understand their topic of study at spatial humanities and yet those looking to spatially index and query unstructured texts. I approached the topic from the perspective of my native language, Finnish, with the aim of introducing a geoparser for the language and, through the creation process, also learn what sort of challenges might emerge when tackling this issue on a language that differs from English in many ways.

The concrete contributions of this work, which are shared openly, are a geoparser for Finnish texts, Finger[23], and two datasets to evaluate Finnish geoparsers[24]. The toponym recognition pipeline of Finger is built on a deep learning language model and the toponyms resolver on a database query – both of these systems, as well as geoparsing on a conceptual level, can be expanded on in many ways. Yet the initial evaluation shows promise: the system performed decently on the two datasets. I hope that, through further development and applied use, geoparsing and Finger will find their place as a source of data in Finnish geospatial research.

Acknowledgements

Suurkiitokset ohjaajilleni Tuuli Toivoselle ja Tuomo Hiippalalle opastuksesta alkuvaiheen ideahämärästä aina viimeiseen poistettuun puhekielisyyteen. Ilman panostanne, neuvojanne ja tukeanne matka olisi ollut paljon takkuisempi ja lopputulos kehnompi. Kiitokset Ainolle ja Emilille annotointitalkoissa puurtamisesta: korpuksista ei olisi tullut mitään ilman sitä. Vielä kiitos Ainolle tuesta, haastamisesta, kirjoitussavotoista ja kaikesta muustakin. Kiitän Aleksandra Elbakjania ja Aaron Swartzia työstään ja uhrauksistaan vapaan tieteen puolesta. Kiitokset geoinformatiikan maisteriseminaarin osallistujille ja järjestäjille kiinnostavista keskusteluista synkimpienkin koronakurimusten aikana. Lopuksi kiitos sinulle, että luit graduni. Tai ainakin selasit loppuun. Kiitti!

[23] https://github.com/Tadusko/fi-geoparser
[24] https://github.com/Tadusko/finger-corpora

Literature

Acheson, E., Sabbata, S. D., & Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, *64*, 309–320. https://doi.org/10.1016/j.compenvurbsys.2017.03.007

Adams, B. (2017). Wāhi, a discrete global grid gazetteer built using linked open data. *International Journal of Digital Earth*, *10*(5), 490–503. https://doi.org/10.1080/17538947.2016.1229819

Ainiala, T. (2018). Paikannimien kerrostumat maaseudulla ja kaupungissa nimistöntutkijan kohteina. *Elore*, *25*(1). https://doi.org/10.30666/elore.72817

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal of Humanities and Arts Computing*, *9*(1), 15–35. https://doi.org/10.3366/ijhac.2015.0136

Alex, B., Grover, C., Tobin, R., & Oberlander, J. (2019). Geoparsing historical and contemporary literary text set in the City of Edinburgh. *Language Resources and Evaluation*, *53*(4), 651–675. https://doi.org/10.1007/s10579-019-09443-x

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging Web content. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 273–280. https://doi.org/10.1145/1008992.1009040

Avvenuti, M., Cresci, S., Vigna, F. D., Fagni, T., & Tesconi, M. (2018). CrisMap: A Big Data Crisis Mapping System Based on Damage Detection and Geoparsing. *Information Systems Frontiers*, *20*(5), 993–1011. https://doi.org/10.1007/s10796-018-9833-z

Barker, E., Simon, R., Isaksen, L., & Cañamares, P. de S. (2016). The Pleiades Gazetteer and the Pelagios Project. In M. L. Berman, R. Mostern, & H. Southall (Eds.), *Placing Names: Enriching and Integrating Gazetteers* (pp. 97–109). Indiana University Press. http://oro.open.ac.uk/48328/

Bender, E. (2019). The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

Blaschke, T., Merschdorf, H., Cabrera-Barona, P., Gao, S., Papadakis, E., & Kovacs-Györi, A. (2018). Place versus Space: From Points, Lines and Polygons in GIS to Place-Based Representations Reflecting Language and Culture. *ISPRS International Journal of Geo-Information*, *7*(11), 452. https://doi.org/10.3390/ijgi7110452

Bol, P. K. (2013). On the Cyberinfrastructure for GIS-Enabled Historiography: Space–Time Integration in Geography and GIScience. *Annals of the Association of American Geographers*, *103*(5), 1087–1092. https://doi.org/10.1080/00045608.2013.792178

Buscaldi, D., & Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, *22*(3), 301–313. https://doi.org/10.1080/13658810701626251

Cardoso, A. B., Martins, B., & Estima, J. (2019). Using Recurrent Neural Networks for Toponym Resolution in Text. In P. Moura Oliveira, P. Novais, & L. P. Reis (Eds.), *Progress in Artificial Intelligence* (Vol. 11805, pp. 769–780). Springer International Publishing. https://doi.org/10.1007/978-3-030-30244-3_63

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

Chang, L.-H. (2021). Towards Bilingually Competent Deep Language Modeling. *Proceedings of the ESSLLI Student Session 2021*.

Cope, A., & Kelso, N. (2015). Who's On First. *Who's On First*. https://whosonfirst.org/blog/2015/08/18/who-s-on-first/

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, *29*, 2382–2388.

DeLozier, G., Wing, B., Baldridge, J., & Nesbit, S. (2016). Creating a Novel Geolocation Corpus from Historical Texts. *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016)*, 188–198. https://doi.org/10.18653/v1/W16-1721

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Du, S., Wang, X., Feng, C.-C., & Zhang, X. (2017). Classifying natural-language spatial relation terms with random forest algorithm. *International Journal of Geographical Information Science*, *31*(3), 542–568. https://doi.org/10.1080/13658816.2016.1212356

Eilander, D., Trambauer, P., Wagemaker, J., & van Loenen, A. (2016). Harvesting Social Media for Generation of Near Real-time Flood Maps. *Procedia Engineering*, *154*, 176–183. https://doi.org/10.1016/j.proeng.2016.07.441

Frank, A., & Mark, D. (1991). Language Issues for Geographical Information Systems. In *Geographic Information Systems: Principles and Applications* (1st ed.). Longman Scientific and Technical.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., & Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, *31*(6), 1245–1271. https://doi.org/10.1080/13658816.2016.1273357

Gelernter, J., & Mushegian, N. (2011). Geo-parsing Messages from Microtext. *Transactions in GIS*, *15*(6), 753–773. https://doi.org/10.1111/j.1467-9671.2011.01294.x

Goodchild, M. F., & Hill, L. L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, *22*(10), 1039–1044. https://doi.org/10.1080/13658810701850497

Gregory, I., Donaldson, C., Murrieta-Flores, P., & Rayson, P. (2015). Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, *9*(1), 1–14. https://doi.org/10.3366/ijhac.2015.0135

Gritta, M. (2019). *Where are you talking about? Advances and challenges of geographic analysis of text with application to disease monitoring* [PhD thesis, University of Cambridge]. https://doi.org/10.17863/CAM.41821

Gritta, M., Pilehvar, M. T., & Collier, N. (2018). Which Melbourne? Augmenting geocoding with maps. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, *1*, 1285–1296. https://doi.org/10.18653/v1/p18-1119

Gritta, M., Pilehvar, M. T., & Collier, N. (2020). A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics. *Language Resources and Evaluation*, *54*(3), 683–712. https://doi.org/10.1007/s10579-019-09475-3

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2017). What's missing in geographical parsing? *Language Resources and Evaluation*, *52*(2), 603–623. https://doi.org/10.1007/s10579-017-9385-8

Hahmann, S., & Burghardt, D. (2013). How much information is geospatially referenced? Networks and cognition. *International Journal of Geographical Information Science*, *27*(6), 1171–1189. https://doi.org/10.1080/13658816.2012.743664

Harley, J. B. (1989). Deconstructing the map. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *26*(2), 1–20. https://doi.org/10.3138/E635-7827-1757-9T53

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., & Ginter, F. (2014). Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation*, *48*(3), 493–531. https://doi.org/10.1007/s10579-013-9244-1

Heikinheimo, V., Di Minin, E., Tenkanen, H., Hausmann, A., Erkkonen, J., & Toivonen, T. (2017). User-Generated Geographic Information for Visitor Monitoring in a National Park: A Comparison of Social Media Data and Visitor Survey. *ISPRS International Journal of Geo-Information*, *6*, 85. https://doi.org/10.3390/ijgi6030085

Heino, E. (2017). *Sotahistorian kuvaaminen ja rikastaminen linkitettynä datana* [Master's Thesis, University of Helsinki]. http://hdl.handle.net/10138/229154

Hiippala, T., Väisänen, T., Toivonen, T., & Järv, O. (2020). Mapping the languages of Twitter in Finland: Richness and diversity in space and time. *Neuphilologische Mitteilungen*, *121*(1), 12–44. https://doi.org/10.51814/nm.99996

Hill, L. L. (2006). *Georeferencing: The geographic associations of information*. MIT Press.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/zenodo.1212303

Hu, Y. (2018a). Geospatial Semantics. In *Comprehensive Geographic Information Systems* (pp. 80–94). Elsevier. https://doi.org/10.1016/B978-0-12-409548-9.09597-X

Hu, Y. (2018b). Geo-text data and data-driven geospatial semantics. *Geography Compass*, *12*(11), e12404. https://doi.org/10.1111/gec3.12404

Hu, Y., & Adams, B. (2021). Harvesting Big Geospatial Data from Natural Language Texts. In *Handbook of Big Geospatial Data* (1st ed., pp. 487–508). Springer Nature.

Hu, Y., Mao, H., & McKenzie, G. (2019). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, *33*(4), 714–738. https://doi.org/10.1080/13658816.2018.1458986

Hulden, M., Silfverberg, M., & Francom, J. (2015). Kernel Density Estimation for Text-Based Geolocation. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 145–150.

Järv, O. (2020). Can we use Twitter data to estimate population distribution in Finland? *Digital Geography Lab Blog*. https://blogs.helsinki.fi/digital-geography/2020/01/12/estimating-finnish-population-from-twitter-data/

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., & McKenzie, G. (2016). Things and Strings: Improving Place Name Disambiguation from Short Texts by Combining Entity Co-Occurrence with Topic Modeling. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge Engineering and Knowledge Management* (Vol. 10024, pp. 353–367). Springer International Publishing. https://doi.org/10.1007/978-3-319-49004-5_23

Jurafsky, D., & Martin, J. (2022). *Speech and Language Processing* (3rd ed.) [Book draft of January 12, 2022]. https://web.stanford.edu/~jurafsky/slp3/

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y., & Ruths, D. (2015). Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. *Proceedings of the*

International AAAI Conference on Web and Social Media, *9*, 188–197. https://ojs.aaai.org/index.php/ICWSM/article/view/14627

Kanerva, J., Ginter, F., & Salakoski, T. (2021). Universal Lemmatizer: A Sequence to Sequence Model for Lemmatizing Universal Dependencies Treebanks. *Natural Language Engineering*, *27*(5), 545–574. https://doi.org/10.1017/S1351324920000224

Karimzadeh, M. (2016). Performance Evaluation Measures for Toponym Resolution. *Proceedings of the 10th Workshop on Geographic Information Retrieval*, 1–2. https://doi.org/10.1145/3003464.3003472

Karimzadeh, M., & MacEachren, A. M. (2019). GeoAnnotator: A Collaborative Semi-Automatic Platform for Constructing Geo-Annotated Text Corpora. *ISPRS International Journal of Geo-Information*, *8*(4). https://doi.org/10.3390/ijgi8040161

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., & Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, *23*(1), 118–136. https://doi.org/10.1111/tgis.12510

Karl, J. W. (2019). Mining location information from life- and earth-sciences studies to facilitate knowledge discovery. *Journal of Librarianship and Information Science*, *51*(4), 1007–1021. https://doi.org/10.1177/0961000618759413

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Koivisto, S. (2021). *Twitter as an Indicator of Sports Activities in the Helsinki Metropolitan Area* [Master's Thesis, University of Helsinki]. http://hdl.handle.net/10138/333268

Kulkarni, S., Jain, S., Hosseini, M. J., Baldridge, J., Ie, E., & Zhang, L. (2020). Spatial Language Representation with Multi-Level Geocoding. *ArXiv:2008.09236 [Cs]*. http://arxiv.org/abs/2008.09236

Laparra, E., & Bethard, S. (2020). A Dataset and Evaluation Framework for Complex Geographical Description Parsing. *Proceedings of the 28th International Conference on Computational Linguistics*, 936–948. https://doi.org/10.18653/v1/2020.coling-main.81

Leidner, J. (2007). *Toponym Resolution in Text* [PhD Thesis, University of Edinburgh]. http://hdl.handle.net/1842/1849

Leidner, J., & Lieberman, M. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, *3*(2), 5–11. https://doi.org/10.1145/2047296.2047298

Lieberman, M. D., Samet, H., & Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 201–212. https://doi.org/10.1109/ICDE.2010.5447903

Luoma, J., Chang, L.-H., Ginter, F., & Pyysalo, S. (2021). Fine-grained Named Entity Annotation for Finnish. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 135–144. https://aclanthology.org/2021.nodalida-main.14

Luoma, J., Oinonen, M., Pyykönen, M., Laippala, V., & Pyysalo, S. (2020). A Broad-coverage Corpus for Finnish Named Entity Recognition. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020* (pp. 4615–4624). European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.567/

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., & Blanford, J. (2011). SensePlace2: GeoTwitter analytics support for situational awareness. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 181–190. https://doi.org/10.1109/VAST.2011.6102456

Matsuda, K., Sasaki, A., Okazaki, N., & Inui, K. (2015). Annotating Geographical Entities on Microblog Text. *Proceedings of The 9th Linguistic Annotation Workshop*, 85–94. https://doi.org/10.3115/v1/W15-1609

Melo, F., & Martins, B. (2017). Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, *21*(1), 3–38. https://doi.org/10.1111/tgis.12212

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, *36*(4), 1–27. https://doi.org/10.1145/3202662

Miller, H., & Goodchild, M. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449–461. https://doi.org/10.1007/s10708-014-9602-6

Moncla, L., Gaio, M., Joliveau, T., Lay, Y.-F. L., Boeglin, N., & Mazagol, P.-O. (2019). Mapping urban fingerprints of odonyms automatically extracted from French novels. *International Journal of Geographical Information Science*, *33*(12), 2477–2497. https://doi.org/10.1080/13658816.2019.1584804

Monteiro, B. R., Davis, C. A., & Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, *96*, 23–34. https://doi.org/10.1016/j.cageo.2016.07.017

Montello, D. R., Friedman, A., & Phillips, D. W. (2014). Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, *28*(9), 1802–1820. https://doi.org/10.1080/13658816.2014.900178

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1659–1666. https://aclanthology.org/L16-1262

Pilehvar, M. T., & Camacho-Collados, J. (2020). *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning* (1st ed.). Morgan & Claypool. https://doi.org/10.2200/S01057ED1V01Y202009HLT047

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends in Information Retrieval*, *12*(2–3), 164–318. http://dx.doi.org/10.1561/1500000034

Purves, R. S., & Derungs, C. (2015). From Space to Place: Place-Based Explorations of Text. *International Journal of Humanities and Arts Computing*, *9*(1), 74–94. https://doi.org/10.3366/ijhac.2015.0139

Reback, J., jbrockmendel, McKinney, W., Bossche, J. V. den, Augspurger, T., Cloud, P., Hawkins, S., gfyoung, Sinhrks, Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Hoefler, P., Naveh, S., Garcia, M., Schendel, J., … Dong, K. (2021). *pandas-dev/pandas: Pandas 1.3.0* (v1.3.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5060318

Ringland, N. (2016). *Structured Named Entities* [PhD Thesis, The University of Sydney]. http://hdl.handle.net/2123/14558

Rocco, L. D., Dassereto, F., Bertolotto, M., Buscaldi, D., Catania, B., & Guerrini, G. (2021). Sherloc: A knowledge-driven algorithm for geolocating microblog messages at sub-city level. *International Journal of Geographical Information Science*, *35*(1), 84–115. https://doi.org/10.1080/13658816.2020.1764003

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Rose-Redwood, R., Alderman, D., & Azaryahu, M. (2010). Geographies of toponymic inscription: New directions in critical place-name studies. *Progress in Human Geography*, *34*(4), 453–470. https://doi.org/10.1177/0309132509351042

Ruokolainen, T., Kauppinen, P., Silfverberg, M., & Lindén, K. (2020). A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, *54*(1), 247–272. https://doi.org/10.1007/s10579-019-09471-7

Santos, J., Anastácio, I., & Martins, B. (2015). Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, *80*(3), 375–392. https://doi.org/10.1007/s10708-014-9553-y

Stock, K., Jones, C. B., Russell, S., Radke, M., Das, P., & Aflaki, N. (2021). Detecting geospatial location descriptions in natural language text. *International Journal of Geographical Information Science*, 1–38. https://doi.org/10.1080/13658816.2021.1987441

Stock, K., & Yousaf, J. (2018). Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. *International Journal of Geographical Information Science*, *32*(6), 1087–1116. https://doi.org/10.1080/13658816.2018.1432861

Summers, E., Brigadir, I., Kemenade, H. van, Hames, S., Binkley, P., tinafigueroa, Ruest, N., Walmir, Chudnov, D., recrm, celeste, Chosak, A., McCain, R. M., Milligan, I., Segerberg, A., Shahrokhian, D., Walsh, M., Lausen, L., Woodward, N., … Kerchner, D. (2021). *DocNow/twarc:* (v2.8.2) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5758827

Tambuscio, M., & Andrews, T. L. (2021). Geolocation and Named Entity Recognition in Ancient Texts: A Case Study about Ghewond's Armenian History. In M. Ehrmann, F. Karsdorp, M. Wevers, T. L. Andrews, M. Burghardt, M. Kestemont, E. Manjavacas, M. Piotrowski, & J. van Zundert (Eds.), *Proceedings of the Conference on Computational Humanities Research, CHR2021, Amsterdam, The Netherlands, November 17-19, 2021* (Vol. 2989, pp. 136–148). CEUR-WS.org. http://ceur-ws.org/Vol-2989/short_paper28.pdf

Tateosian, L., Guenter, R., Yang, Y.-P., & Ristaino, J. (2017). Tracking 19th Century Late Blight from Archival Documents using Text Analytics and Geoparsing. *Free and Open Source Software for Geospatial (FOSS4G): Conference Proceedings*, *17*, 16. https://doi.org/10.7275/R5J964K5

Tkachenko, M., Malyuk, M., Shevchenko, N., Holmanyuk, A., & Liubimov, N. (2020). *Label Studio: Data labeling software*. https://github.com/heartexlabs/label-studio

Tobin, R., Grover, C., Byrne, K., Reid, J., & Walsh, J. (2010). Evaluation of Georeferencing. *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR'10)*.

Tuan, Y.-F. (1979). Space and Place: Humanistic Perspective. In S. Gale & G. Olsson (Eds.), *Philosophy in Geography* (pp. 387–427). Springer Netherlands. https://doi.org/10.1007/978-94-009-9394-5_19

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *ArXiv Preprint ArXiv:1912.07076*.

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, *32*(1), 1–29. https://doi.org/10.1080/13658816.2017.1368523

Wang, J., & Hu, Y. (2019a). Are We There yet? Evaluating State-of-the-Art Neural Network Based Geoparsers Using EUPEG as a Benchmarking Platform. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*, 1–6. https://doi.org/10.1145/3356991.3365470

Wang, J., & Hu, Y. (2019b). Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, *23*(6), 1393–1419. https://doi.org/10.1111/tgis.12579

Wang, J., Hu, Y., & Joseph, K. (2020). NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS*, *24*(3), 719–735. https://doi.org/10.1111/tgis.12627

Wang, X., Ma, C., Zheng, H., Liu, C., Xie, P., Li, L., & Si, L. (2019). DM_NLP at SemEval-2018 Task 12: A Pipeline System for Toponym Resolution. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 917–923. https://doi.org/10.18653/v1/S19-2156

Yan, B., Mai, G., Hu, Y., & Janowicz, K. (2021). Harnessing Heterogeneous Big Geospatial Data. In *Handbook of Big Geospatial Data*. Springer.

Zheng, X., Han, J., & Sun, A. (2018). A Survey of Location Prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), 1652–1671. https://doi.org/10.1109/TKDE.2018.2807840

Appendix A. Annotation practices for Finger-news and Finger-tweets

Below is a brief description of the nuances of the annotation process, especially how unclear or edge-cases were handled.

Annotating toponyms

- We decided against marking toponyms as noun modifiers (e.g. Suomen pääministeri, *the Finnish Prime minister*). There might be some unevenness in the annotation of these.
    - Ranskan vallankumous (*the French revolution*): seen as an event, *Ranska* not marked
- We tried to assess from the context whether the toponym was spatial or referred to an organization etc. There might be some unevenness on these cases.

- Place name embedded in a larger hashtags were not marked (#salonkaupunki, #EspooLiikkuu) whereas simpler hashtags (#espoo) were. The hashtag sign is not included in the span.
    - If the word was conjugated, often the root was separated from the affix via an apostrophe. The whole word was marked in these cases (#Espoo'ssa)
    - From "Vantaa-Keravan hyvinvointialueen" the highlighted was annotated
- Vague descriptors (e.g. *liepeillä* in Helsingin liepeillä, *in the vicinity of Helsinki*) were **not** included in the spans. Established ones are (e.g. Turun saaristo, *the Turku archipelago*).
- Tweets are sometimes mixed-language. Toponyms are marked nonetheless (e.g. the hashtag #borgå)

Annotating locations

- The annotator looked up WGS 84 coordinates from GeoNames' web service and copied them (latitude-longitude format).
    - If the toponym was not in GeoNames, the annotator looked up NLS Place Names through NimiSampo, OpenStreetMap and Google Maps.
    - Sometimes, especially in the case of facilities, the annotator had to approximate the location visually so that it is, e.g., at the center of the building.
- In the case of wholly unavailable toponyms (e.g. the defunct administrative area of Kaakkois-Pirkanmaa), we decided to keep the annotations but replace the coordinates with a NaN tag.
- A common occurrence especially for Finnish municipalities was that there were multiple options for them: one a "seat" of the municipality and the other for the "third-order administrative division". We interpreted these so that the first refers to e.g. the central town of the municipality (*taajama*, *keskus*) and the second to the area as a whole. Therefore, the latter option is used in most of the cases, unless a reference to the administrative center was clear from the context.