

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/382850674>

# Geographical and linguistic perspectives on developing geoparsers with generic resources

Article in *International Journal of Geographical Information Science* · June 2024

DOI: 10.1080/13658816.2024.2369539

CITATIONS

2

READS

43

3 authors:



**Tatu Leppämäki**

University of Helsinki

1 PUBLICATION 2 CITATIONS

SEE PROFILE



**Tuuli Toivonen**

University of Helsinki

151 PUBLICATIONS 6,688 CITATIONS

SEE PROFILE

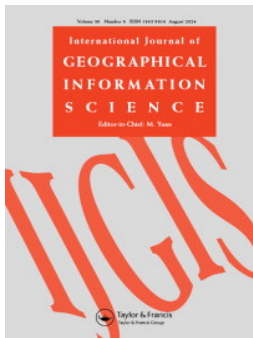


**Tuomo Hiippala**

University of Helsinki

65 PUBLICATIONS 2,100 CITATIONS

SEE PROFILE



## Geographical and linguistic perspectives on developing geoparsers with generic resources

Tatu Leppämäki, Tuuli Toivonen & Tuomo Hiippala

**To cite this article:** Tatu Leppämäki, Tuuli Toivonen & Tuomo Hiippala (30 Jun 2024): Geographical and linguistic perspectives on developing geoparsers with generic resources, International Journal of Geographical Information Science, DOI: [10.1080/13658816.2024.2369539](https://doi.org/10.1080/13658816.2024.2369539)

**To link to this article:** <https://doi.org/10.1080/13658816.2024.2369539>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 30 Jun 2024.



[Submit your article to this journal](#)



Article views: 401



[View related articles](#)



[View Crossmark data](#)



RESEARCH ARTICLE



OPEN ACCESS



# Geographical and linguistic perspectives on developing geoparsers with generic resources

Tatu Leppämäki<sup>a</sup> , Tuuli Toivonen<sup>a</sup> and Tuomo Hiippala<sup>a,b</sup>

<sup>a</sup>Digital Geography Lab, Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland; <sup>b</sup>Department of Languages, University of Helsinki, Helsinki, Finland

## ABSTRACT

Geoparsers aim to find place names in unstructured texts and locate them geographically. This process produces georeferenced data usable for spatial analyses or visualisations. Much geoparsing research and development has thus far focused on the English language, yet languages are not alike. Geoparsing them may necessitate language-specific processing steps or data for training geoparsing systems. In this article, we applied generic language and GIS resources to geoparsing Finnish texts. We argue that using generic resources can ease the development of geoparsers, and free up resources to other tasks, such as annotating evaluation corpora. A quantitative evaluation on new human-annotated news and tweet corpora indicates robust overall performance. A systematic analysis of the geoparser output reveals errors and their causes at each processing step. Some of the causes are specific to Finnish, and offer insights to geoparsing other morphologically complex languages as well. Our results highlight how the language of the input text affects geoparsing. Additionally, we argue that toponym resolution metrics based on error distance have limitations, and proposed metrics based on spatial intersection with ground-truth polygons.

## ARTICLE HISTORY

Received 4 November 2023  
Accepted 14 June 2024

## KEYWORDS

Geoparsing; open source; toponym recognition; toponym resolution; geoparsing evaluation

## 1. Introduction

Geographically referenced texts, or *geo-text data* (Y. Hu 2018), can be used to study a wide range of geographical phenomena. Although much data is claimed to be geographically referenced (Hahmann and Burghardt 2013), these references may be unstructured, as exemplified by place names in free-form text (Y. Hu 2018), and thus ill-suited to common methods of spatial analysis. Methods for spatially analysing non-coordinate geographical references remain underdeveloped (Purves *et al.* 2019, Janowicz *et al.* 2022) which is why linguistic references to geographical entities must be transformed into geographical coordinates (Y. Hu and Adams 2021). This process is

**CONTACT** Tatu Leppämäki [tatu.leppamaki@helsinki.fi](mailto:tatu.leppamaki@helsinki.fi)

Supplemental data for this article can be accessed online at <https://doi.org/10.1080/13658816.2024.2369539>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

called *geoparsing*. Conceptual and methodological research on geoparsing is actively pursued in fields such as language technology (Gritta *et al.* 2020) and geographical information retrieval (Purves *et al.* 2018). This has led to the development of geoparsers (Tobin *et al.* 2010, Karimzadeh *et al.* 2019, Halterman 2023), geoparsing evaluation corpora (Lieberman *et al.* 2010, Gritta *et al.* 2018b, Wallgrün *et al.* 2018) and standard evaluation methods (Wang and Hu 2019b, Gritta *et al.* 2020).

However, geoparsing systems and resources commonly focus on English or one of its large, close relatives in the Indo-European language family. This type of bias towards English has been recognised, for example, in natural language processing (NLP) (Bender 2011, 2019) and cognitive science (Blasi *et al.* 2022). Languages differ in terms of structure and the amount of geo-text data available for training and testing geoparsers. Just how language affects and limits geoparsing and the development of geoparsers has rarely been discussed.

Access to geoparsers for many languages is necessary to enable wider use of geo-text data, but few such systems exist, and none for the language we focus on, Finnish. Developing geoparsers may require significant investments of time, expertise and data. However, tools and resources available for general NLP and GIS tasks, such as large language models (Devlin *et al.* 2019) and open-source geocoders, offer an alternative for developing methods specifically for geoparsing. In this article, we used such generic tools to geoparse Finnish texts.

Geoparser output must be thoroughly evaluated. However, the standard evaluation metrics for toponym resolution (e.g. Wang and Hu 2019b, Gritta *et al.* 2020) might fail, for example, because of spatial misalignment unrelated to geoparsing. Additionally, the metrics apply only to toponym recognition or resolution individually. Understanding errors at each geoparsing step, and how these affect the subsequent tasks, is necessary if geoparsers are to be further improved (Acheson and Purves 2021). We proposed an alternative toponym resolution method and performed a comprehensive error analysis of our system.

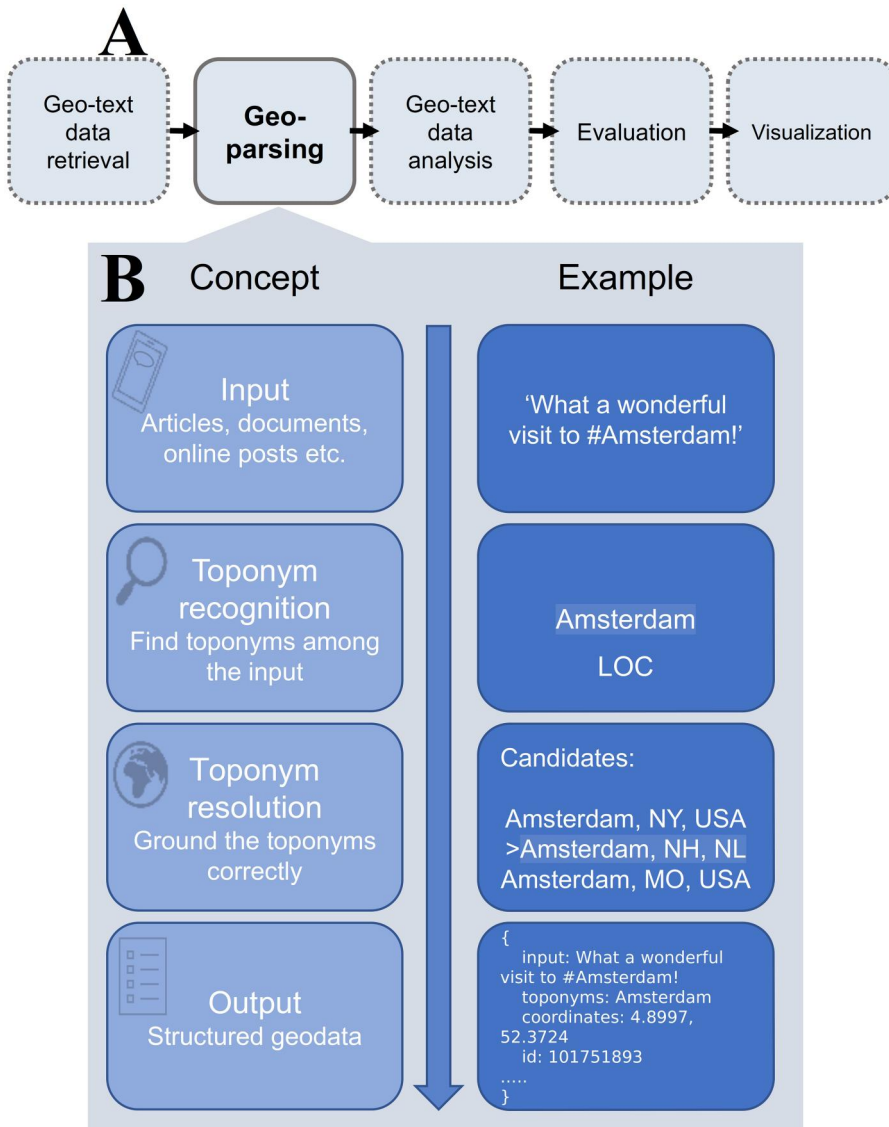
To summarise our contributions:

1. We used generic resources to develop a geoparser for Finnish texts. We then performed a systematic error analysis of the full geoparsing pipeline and identified causes of error at each processing step. Based on the results, we discuss how geoparsing is affected by the target language, and specifically, our experience of applying geoparsing to a morphologically complex language.
2. We published two human-annotated Finnish corpora for tweets and news articles for evaluating geoparsing performance.
3. We proposed two novel toponym resolution metrics (ACC@POLY, ACC@BBOX) based on whether the system's prediction intersects with a ground-truth area.

## 2. Background

### 2.1. Overview of geoparsing

Geoparsing refers to recognising and unambiguously grounding toponyms found in unstructured texts (Gritta *et al.* 2018a, Y. Hu and Adams 2021). This task is generally split into two steps: (1) toponym recognition (sometimes geotagging), which refers to



**Figure 1.** (A) Geoparsing as part of an analysis pipeline for geo-text data, adapted from (Y. Hu 2018, p. 11). (B) An overview of geoparsing steps for an English sentence. In this article, we develop a similar analysis pipeline for sentences fully in Finnish.

identifying references to geographical locations in texts, whereas (2) toponym resolution (sometimes geocoding) refers to unambiguously locating each toponym in a coordinate system. Some also use the term geoparsing to refer to what is here called toponym recognition (e.g. Leidner and Lieberman 2011, Purves *et al.* 2018).

Geoparsers output geographic information in a structured data format (Figure 1(B)). Y. Hu (2018) proposed a workflow for processing texts with spatial linkages, or *geo-text data*, which positions geoparsing as a step for structuring implicit references to geographical information. The structured data can then be applied in subsequent tasks, such as spatio-temporal analyses and data visualisation (Figure 1(A)). This article

focuses exclusively on geoparsing, but we consider the pipeline in Figure 1(A) to be a key motivation for developing geoparsing methods.

Geoparsing is used in diverse disciplines that work with textual data. A recent review by X. Hu *et al.* (2023b) identified seven application domains, ranging from digital humanities (e.g. Gregory *et al.* 2015, Ardanuy *et al.* 2023), to geographical information retrieval (Purves *et al.* 2018). Another domain is the analysis of content on social media platforms, which contain massive volumes of unstructured text. Geoparsed geo-text data from Twitter<sup>1</sup> and other social media platforms have been proposed as sources of information for monitoring and managing rapidly developing events, such as disasters, or events that emit weak signals, such as the spread of diseases (Wang *et al.* 2020, X. Hu *et al.* 2023b). For example, geoparsing is a crucial step in gathering, structuring and visualising disaster information posted online (Avvenuti *et al.* 2018), whereas geoparsed tweets have been used to map informal exercising outside sports venues (Liu *et al.* 2022a).

Both custom-built and generic tools have been applied to geoparsing. By generic tools, we mean existing resources for tasks adjacent to toponym recognition and resolution, which may be applied to geoparsing with little adjustment. For example, toponym recognition has been likened to a special case of named entity recognition (NER) (Leidner and Lieberman 2011, Gritta *et al.* 2020), a NLP task in which spans of text are classified into predetermined categories, such as persons, organisations and locations. Geotxt (Karimzadeh *et al.* 2019) and Mordecai (Halterman 2023) are examples of geoparsers that use generic NER models for toponym recognition.

Toponym recognition methods may be based on rules, gazetteer matching, or machine learning models or their combination (Leidner and Lieberman 2011, X. Hu *et al.* 2023b). Edinburgh geoparser (Tobin *et al.* 2010) is an example of a geoparser that uses hand-crafted symbolic rules, whereas *geoparsepy* adopts a gazetteer matching approach (Middleton *et al.* 2018). A review by X. Hu *et al.* (2023b) introduces toponym recognition methods and example systems in more detail. Recently, large language models, such as the Transformer-based BERT (Devlin *et al.* 2019) and its variants have been increasingly applied to toponym recognition (see e.g. Berragan *et al.* 2022, Ma *et al.* 2022). An advantage of these pretrained models is that they may be further fine-tuned for specific downstream tasks, such as toponym recognition, using relatively small labelled datasets.

Toponym resolution, sometimes referred to as geocoding, involves unambiguously grounding the toponyms recognised in the previous step. Toponym resolution methods often rely on gazetteers, such as OpenStreetMap (OSM)<sup>2</sup>, Who's On First (WoF)<sup>3</sup> and GeoNames<sup>4</sup>, of which GeoNames is the most often used source gazetteer (Gritta *et al.* 2020). Gazetteer-based resolvers rank ambiguous toponyms. In most cases, simple ranking by prominence or population achieves robust baseline performance (Karimzadeh *et al.* 2019, Wang and Hu 2019a). Mordecai 3 uses another approach, a neural network model to re-rank gazetteer responses (Halterman 2023). Gazetteers may have omissions, and have been shown to have inconsistent coverage spatially and across feature types (Acheson *et al.* 2017). An alternative approach is to train models that use linguistic features to predict the most likely location on a local or global grid (Hulden *et al.* 2015, Gritta *et al.* 2018b, Fize *et al.* 2021, Kulkarni *et al.* 2021).

Although such models have achieved impressive results in performance comparisons of toponym resolvers (Wang and Hu 2019a), we are not aware of end-to-end geoparsers that have implemented them.

Toponym recognition and resolution are complicated by reference ambiguity (Amitay *et al.* 2004, Moncla *et al.* 2014). We adopted the term geo/non-geo ambiguity for words with non-geographic meanings (e.g. Paris as a place and a given name), and geo/geo ambiguity when multiple places share the same name (Amitay *et al.* 2004). Linguistic ambiguity highlights that geoparsing is a linguistic as well as geospatial problem – therefore, the ways in which the language processed impacts geoparsing warrants a closer examination.

## 2.2. Linguistic issues in geoparsing

Bender (2019) argues that the language that is being processed, which we call target language, warrants attention in all NLP tasks. Methods developed for languages such as English are not necessarily language-independent, but are affected by the characteristics of the target language, such as its morphological features and the availability of (labelled) data for training and testing systems (Bender 2011, 2019).

Geoparsing is affected by the linguistic features of the target language. One such example is word inflection. English exhibits relatively little inflectional morphology, while other languages modify base words with affixes, producing a wide range of surface forms (Bender 2019). For example, the expression ‘from Helsinki’, which English construes using the preposition *from*, is expressed in Finnish as ‘Helsingistä’ (Helsinki.ELATIVE). Morphological complexity impacts toponym resolution: gazetteers typically store the toponym in the nominative case (e.g. ‘Helsinki’), which is why the surface forms must be transformed to base forms or lemmas for effective querying and matching (Pouliquen *et al.* 2004). This process is known as lemmatisation. As we show below, lemmatisation is non-trivial and is a consistent source of errors in the geoparsing pipeline. Further examples are given in Dewandaru *et al.* (2020), who describe the problem of separating morphologically identical demonyms from toponyms in a geoparsing corpus for Indonesian, whereas Ma *et al.* (2022) observe that toponym recognition in Chinese requires specific considerations due to the lack of separators and capitalisation.

The availability of resources limits the opportunities for training and testing geoparsers. State-of-the-art deep learning models applied to toponym recognition (Wang *et al.* 2020) and deep learning toponym resolvers (e.g. Gritta *et al.* 2018b, Kulkarni *et al.* 2021) require extensive geo-text data to train. Wikipedia, the most common source of this type of geo-text data, is imbalanced between languages in terms of article quantity, quality and metadata coverage. For example, Liu *et al.* (2022a) could not replicate the procedure for generating training data from Wikipedia proposed in Wang *et al.* (2020), due to the lack of hyperlinks and location information in the Finnish Wikipedia.

Corpora for evaluating the performance of geoparsers is another crucial language-specific resource with imbalanced availability. Major languages, prominently English, are well covered (see, e.g. EUPEG by Wang and Hu 2019b) and geo-text data

availability enables creating corpora for these languages programmatically (Gritta *et al.* 2018a). Moreover, there appears to be a link between language and geographical coverage – Wallgrün *et al.* (2018) noted that the majority of toponyms in their Twitter-based corpus GeoCorpora were within countries in the English-speaking world. Our analyses show that this holds true for four other global English corpora, and roughly half of toponyms in these corpora fall within the United States (Supplementary material 1).

Based on the reasons outlined above, extending geoparser coverage to a new language may require processing steps specific to that language and investing significant resources. To our knowledge, few previous studies have applied geoparsing to Finnish, our target language (Pouliquen *et al.* 2004, Liu *et al.* 2022a). We are not aware of previously published end-to-end geoparsers for Finnish texts, nor of any publicly available evaluation corpora for the language.

### 3. Methods

#### 3.1. System overview

To enable the geoparsing of Finnish texts, we developed a system with existing GIS and NLP resources. Our aim was simple development, installation, usage and extension, which in turn free resources elsewhere. The geoparser consists of two components: (1) a NLP component for toponym recognition and (2) a toponym resolver component. The NLP tasks, which include lemmatisation and toponym recognition, are processed by a custom *spaCy* pipeline (Montani *et al.* 2023), as described in Section 3.2. Toponym resolution is performed by querying Pelias geocoder web service, which indexes data from three gazetteers (Section 3.3). The components are wrapped by our geoparser, and the system is evaluated with two manually annotated corpora created for this study (see Sections 4.1–4.3). Figure 2 presents an overview of the system design and workflow of this study, whereas Figure 3 shows an example of

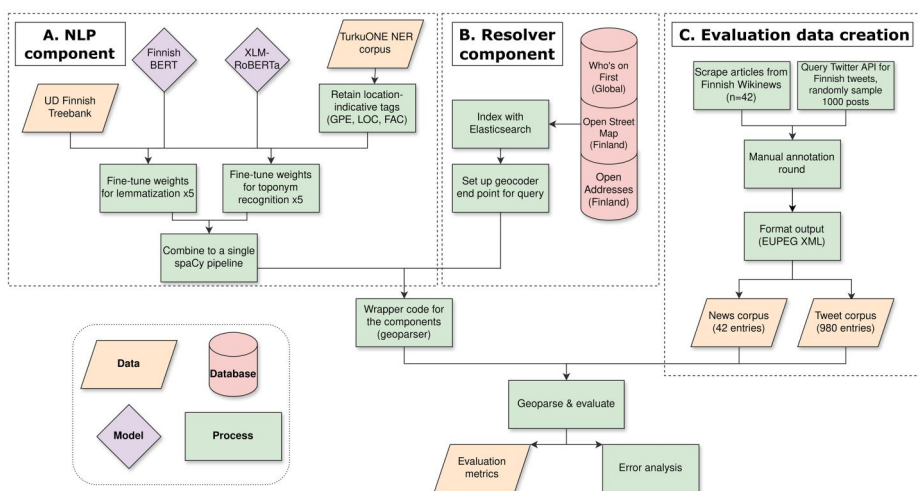
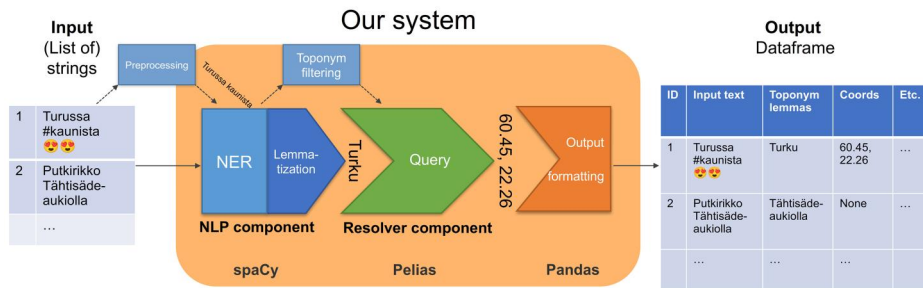


Figure 2. Overview of system design and research process.





**Figure 3.** The proposed system at runtime. The system accepts a list of strings, roughly the length of a sentence or a paragraph, as input, and outputs a dataframe with many columns. Two examples are presented: (1) geoparsed successfully and (2) failed, due to incorrect lemmatisation.

the system at runtime. We published our geoparser as an open-source Python package.

### 3.2. NLP component for lemmatisation and toponym recognition

The NLP component is used to recognise and lemmatise toponyms. We chose the spaCy NLP library (Montani *et al.* 2023) for this purpose, because it implements state-of-the-art methods and has been successfully used in previous geoparsers (e.g. Halterman 2023).

Considering the investments required for developing a rule-based toponym classifier and gazetteer-matching's vulnerability to geo/non-geo ambiguity, we based our NLP component on Transformer-based language models, which have achieved state-of-the-art performance for many linguistic tasks (Devlin *et al.* 2019, Virtanen *et al.* 2019). We used the NER and lemmatisation components available in spaCy, which can use contextual word embeddings from language models as input features. Currently, there are two Transformer-based word embedding models pre-trained with exclusively Finnish input data: Finnish BERT (Virtanen *et al.* 2019) and Finnish RoBERTa.<sup>5</sup> In addition, some multilingual models have been pre-trained on texts including Finnish: mBERT (Devlin *et al.* 2019) and XLM-RoBERTa (Conneau *et al.* 2020). Previous research has noted the inferior performance of multilingual language models compared to monolingual Finnish models for NER (Virtanen *et al.* 2019) and toponym recognition (Liu *et al.* 2022a). We chose Finnish BERT and XLM-RoBERTa to identify differences between monolingual and multilingual models.

We fine-tuned the Transformer-based models to prepare them for the downstream tasks: lemmatisation and toponym recognition. Fine-tuning is a type of transfer learning in which pre-trained model weights are updated through a training process with task-specific data. As there are no large, publicly available toponym recognition corpora for Finnish, we fine-tuned the pre-trained models using existing resources.

For lemmatisation, we used UD Finnish Treebank (Pyysalo *et al.* 2015). The treebank contains about 200k tokens, whose base forms or lemmas have been annotated manually. For toponym recognition, we used the TurkuONE NER corpus (Luoma *et al.* 2021) which contains about 50,000 tokens of Finnish texts from multiple domains,

including news and Wikipedia articles. This corpus follows the OntoNotes (Weischedel *et al.* 2013) schema for named entities, which covers 18 entity types. Three of them are broadly relevant for geoparsing: GPE (geo-political entity; e.g. countries and cities), LOC (location; e.g. mountains and water bodies) and FAC (facility; e.g. buildings, streets) (Luoma *et al.* 2021). We retained only the GPE, LOC and FAC tags, and trained our toponym recogniser on the modified corpus. Train, development and test sets provided for UD Finnish Treebank<sup>6</sup> and TurkuONE<sup>7</sup> were used.

The *spaCy* library allows the creation of NLP pipelines in which consecutive processing steps are applied to the input text. We created a pipeline with the EntityRecognizer and EditTreeLemmatizer components. A single NVIDIA Volta V100 GPU was used for fine-tuning and inference. The training details and hyperparameters are defined in a configuration file which we share for replication. To mitigate the effects of the randomness of the stochastic training process, we fine-tuned each model five times with different random seeds. Average and standard deviation of performance for these models are reported. The models were not fine-tuned jointly for the two tasks, since they use different training corpora. Thus, the final pipeline has two Transformer components.

A strength of this implementation is that the NLP pipeline can be replaced with any *spaCy* pipeline that includes NER and lemmatisation components. A user could follow the fine-tuning procedure outlined above for a custom pipeline, or use a pre-trained pipeline offered by *spaCy*, which currently exist for 24 languages.<sup>8</sup> *spaCy* also offers pre-trained models for Finnish that rely on less computationally intensive word embeddings. These models are trained on the same corpora as our fine-tuned models, which allows using them as baselines. We used small and large variants of *spaCy*'s models for Finnish as baselines in our evaluation.

### 3.3. Resolver component

The geoparser's toponym resolution approach is based on Pelias, an open-source data agnostic geocoder that supports importing gazetteer data from OpenStreetMap (OSM), GeoNames, OpenAddresses and Who's on First (WoF), as well as from plain text files with comma-separated values. The data are indexed with Elasticsearch, which powers the query engine. Matching search results are ranked by Pelias' default method, which ranks locations based on population and prominence in the administrative hierarchy. We chose a gazetteer query and prominence ranking due to their solid performance in previous work (e.g. Karimzadeh *et al.* 2019), although we acknowledge this method is poorly suited for solving geo-geo ambiguity.

By default, our geoparser accepts the first-ranked result as the resolved location. The user may improve the results by providing parameters that limit the query to a certain area or boost results closer to a provided point. Query responses contain metadata related to the toponym in the form of coordinates and bounding boxes, administrative levels above the toponym and unique identifiers; the user has control over what metadata to include in the geoparser's output.

We set up a Pelias instance in a cloud environment using the premade Docker images and installation scripts. We offer a public API endpoint for users. The instance

is initialised with WoF data with global coverage, and OSM and OpenAddresses data with coverage for Finland. WoF provides administrative areas down to the level of neighbourhoods. OSM contains granular data down to the level of facilities and points-of-interest, whereas OpenAddresses provides coverage of street addresses. We recognise that uneven coverage will bias toponym resolution results for Finland but opted for this approach because of the computational resources necessary for indexing and serving the global OSM and OpenAddresses datasets. In addition, we assume that in any case, full coverage is mostly beneficial in Finland, where most native Finnish speakers reside.

## 4. Performance evaluation

### 4.1. Finnish evaluation corpora

To evaluate the performance of the geoparser, we created two manually annotated geoparsing corpora for news and tweets. The first corpus consists of 42 news articles published in the Finnish edition of Wikinews in 2011. The second corpus contains 980 Finnish tweets acquired through Twitter's API in August 2021. The key features of both corpora are given in Table 1. News articles and tweets were chosen to represent linguistic variation across different domains. For example, online posts may contain emojis, abbreviations and informal language use usually absent from news texts.

Toponyms have been defined in various ways. For example, some geoparsing corpora exclude landscape features, whereas others do not include fine-grained toponyms (Wang and Hu 2019b, p. 11). To reflect the varying annotation approaches found in the literature, we adopted distinct toponym definitions from two English corpora for each of our corpora: only administrative units (countries, regions, cities, etc.) were annotated in the news corpus, similar to the GeoVirus corpus of disease-related news articles (Gritta *et al.* 2018b). A more expansive schema was chosen for the tweet corpus, which involved annotating fine-grained natural features, facilities and street addresses, similar to GeoCorpora (Wallgrün *et al.* 2018).

Each toponym span was manually annotated with a LOC token by the first author or one of two volunteers using annotation software Label Studio (Tkachenko *et al.* 2020). Location information attached to each toponym was acquired from GeoNames. Early evaluation revealed systematic errors. For example, although the geoparser correctly resolved the coordinate point for 'Finland', distance between the prediction and ground-truth coordinate was nonetheless 105 km. The errors were caused by spatial misalignment between locations in GeoNames and the gazetteer data used by the

**Table 1.** Information about the evaluation corpora.

Dataset	Documents	Tokens	Toponyms	Mean toponyms/document	Time period	Location types
News	42	6352	189	4.5	2011	Countries, continents: 55% Regions, cities: 45%
Tweets	980	22,513	498	0.51	8/2021	Regions, cities: 49% Countries, continents: 37% POIs, facilities: 9.5% Sub-city regions: 2.9% Waterbodies: 1.9%

The classification of toponyms is based on the *placetype* attribute in the Who's on First gazetteer.

resolver component. Therefore, a second annotation round was deemed necessary for a fair evaluation. The first author examined each toponym and added a gazetteer id, coordinate point, polygon and bounding box information from the Pelias instance described in [Section 3.3](#). About 5% of the locations did not have polygonal representations. In this case, dummy bounding boxes of  $1 \text{ km}^2$  were created around the point coordinate. The mean distance between GeoNames coordinates and the new coordinate points was 146 km in the news and 82 km in the tweet dataset. To evaluate lemmatisation accuracy, each toponym was also paired with its lemma. Details of the annotation process are provided in Leppämäki (2022). The corpora are shared openly alongside the geoparser presented above.

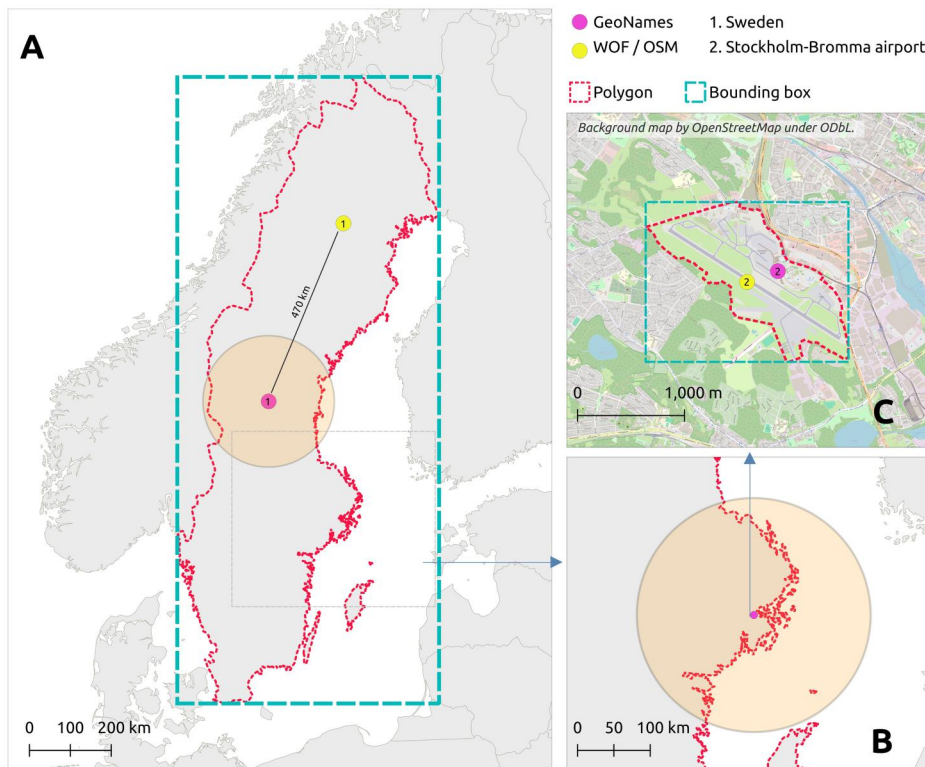
## 4.2. Error metrics

We evaluated the performance of our geoparser with toponym recognition and resolution metrics used in previous work (see e.g. Wang and Hu 2019b, Gritta *et al.* 2020). To evaluate the performance for toponym recognition, we compared the predicted toponym spans with the gold standard corpus, and calculated precision, recall, and F1-scores, which are widely used for evaluating performance for NER. Precision indicates how many correct toponyms the system has found: it is the number of correctly recognised toponyms, true positives, divided by the number of all recognised toponyms including false positives. Recall indicates how many toponyms the system missed: it is the ratio of true positives divided by all toponyms, including the ones the geoparser should have found. F-score balances these two measures – it is the harmonic mean of precision and recall. Finally, we report on lemmatisation accuracy, the proportion of correctly lemmatised toponyms.

Toponym resolution evaluation metrics are commonly based on measures of error distance, that is, the distance between predicted and actual location in kilometres (Wang and Hu 2019b, Gritta *et al.* 2020). Various descriptive statistics may be derived from the error distances: common choices include mean and median. Accuracy@k is the proportion of predictions falling within a set distance from the ground-truth location. By convention, this value is 161 km (100 miles), which is why the metric is called ACC@161. Mean distance, median distance and ACC@161 are global metrics that do not consider the distribution of error distances and treat every error as equally significant. In contrast, the area under the curve (AUC) measures how large the error distances are, compared to the largest possible error distances, half the circumference of Earth (Gritta *et al.* 2020). Following the recommendation in Gritta *et al.* (2020), we have reported the proportion of successfully resolved toponyms. Toponym resolution results are affected by the errors made in previous processing steps. To evaluate the toponym resolver on its own merits, we reported the performance of the system in an optimal case, in which all the toponyms were correctly recognised and lemmatised.

### 4.2.1. ACC@POLY and ACC@BBOX

Metrics based on error distance do not measure performance accurately in all cases – the exact location of a point representing an area is arbitrary, yet that location is used to calculate the error distance. First, predicted and ground-truth coordinate points may



**Figure 4.** Intuition behind using polygon intersection for evaluating toponym resolution. Two example toponyms ('Sweden' and 'Stockholm-Bromma airport') are represented by polygon boundaries, bounding boxes and coordinate points from GeoNames and Who's on First/OpenStreetMap gazetteers. The orange 161-km buffers around points from GeoNames represent the area used for evaluating  $ACC@161$ . (A) Error distance for two valid representations of 'Sweden' is over 470 km apart. One hundred and sixty-one kilometres buffer is not lenient enough, whereas the polygons cover both. (B) A buffer of 161 km is overly allowing, and covers the whole Stockholm region (C) instead of only the area of the airport.

originate from different sources and be spatially misaligned. Evaluation in this case is not like-for-like (Gritta *et al.* 2020). Second, the significance of an error distance is dependent on the scale of the location: an error of 10 km may be considerable for a coffee shop but is insignificant for a country.  $Accuracy@161$  is less susceptible to misalignment, but the constant error distance limit is not optimal, since 161 km is too forgiving for granular locations, and too strict for coarse locations (see Figure 4). A metric that measures performance on locations of various scales is needed (Gritta *et al.* 2018a).

We propose using polygon boundaries of ground-truth locations as a solution that works for multiple scales and is less sensitive to spatial misalignment of coordinate points. Using the same approach as the  $ACC@k$  metric, we propose calculating the proportion of coordinate predictions resolved within a ground-truth area. We report this metric as accuracy at polygon, or  $ACC@POLY$ . Storing accurate polygon geometries, especially for large areas, takes a significant amount of storage space. Additionally, polygon geometries may not be readily available. For these reasons, we additionally propose using minimum bounding boxes of ground-truth areas:  $ACC@BBOX$ . The caveat of this

approach is that bounding boxes are necessarily at least as large as the polygon boundaries and include areas outside of them. The operation of these metrics is demonstrated in Figure 4.

### 4.3. Error analysis methods

In addition to reporting evaluation metrics, error analyses have been performed to better understand the causes and outcomes of errors (e.g. Gelernter and Mushegian 2011, Acheson and Purves 2021). We manually analysed the output of our geoparser using the best performing toponym recognition model. We approached this analysis from the perspective of a user running a geo-text data pipeline (Figure 1). Considering the adage garbage-in, garbage-out, it is relevant to understand how much ‘noise’ is generated by the geoparser, and how much information is lost. Multiple processing steps may introduce errors, whose causes need to be understood. Both corpora were processed by the geoparser and the first author systematically evaluated the outputs.

Each toponym was classified into one of three outcomes:

- *Geo-text data*: A toponym is correctly recognised and located. Overlap with the ground-truth bounding box is considered a correct resolution outcome.
- *Information loss*: A toponym that should have been recognised and located and was not. Thus, information was lost.
- *Noise*: A toponym is falsely located.

Toponym recognition may result in true positive, false positive and false negative; similarly, a toponym may be correctly resolved, falsely resolved or not resolved at all. There are multiple causes of failure for these steps, which the annotator attempted to identify. The causes identified, and their descriptions are presented in Table 2. The causes are not exclusive, that is, a toponym may be falsely recognised, lemmatised and resolved. However, for simplicity, each toponym was given only one error cause based on which error happened at the earliest processing step, following the order of toponym recognition, lemmatisation and toponym resolution.

## 5. Results

### 5.1. Performance evaluation results

Table 3 reports the toponym recognition evaluation results. The Transformer-based models outperformed the baseline models for both corpora and all metrics. The Transformer-based models performed equally well on the news dataset. Differences arose on the larger and more challenging tweet corpus, in which the monolingual model outperformed the multilingual one. Lower recall indicates the fine-tuned XLM-RoBERTa failed to recognise some of the toponyms found by the monolingual model. Models initialised with different random seeds vary, especially on recall, as shown by high standard deviations. The multilingual model achieves the highest lemmatisation accuracy on both corpora.

**Table 2.** Identified toponym recognition, lemmatisation and resolution errors and their causes.

Processing step	Error type	Error cause	Error cause description	Example	Possible outcomes
Toponym resolution	False positive	<i>Actual false positive</i>	The prediction is not a toponym.	Event: <i>Helsinki Half Marathon</i> ; Organisation: <i>Sämediggi Kirstinkatu-; Keski-Savo seudun (Keski-Savo region, should be: Keski-Savo).</i>	Noise/unresolved
		<i>Toponym boundary error</i>	The prediction does not contain exactly the same characters as the ground-truth toponym. The difference can be as minor as a single character.		Geo-text data/noise
		<i>Annotation error</i>	The prediction, although missing from the corpus, is accepted as a valid toponym. Thus, there has been an error in the annotation process.	OMBRA, an unannotated point-of-interest	Geo-text data
Lemmatisation	False negative	<i>Schema mismatch</i>	The toponym recognition system and the corpus use different schemas.	Water bodies in the news corpus are not annotated but are recognised.	Noise
		<i>True false negative</i>	Toponym recogniser failed to predict a toponym that is present in the corpus.	<i>Suez, Akaa</i>	Information loss
		<i>Lemmatisation error</i>	The predicted lemma does not match the ground-truth lemma.	<i>Uudenmata (Uusimaa), Tapiola urheilupuisto (Tapiola sports park)</i>	Information loss/noise
Toponym resolution	Resolved	<i>Geo-geo error</i>	Resolved to wrong entry due to reference ambiguity.	<i>Pallas resolved to Pallas, Austria instead of the Pallas fell in Finland.</i>	Noise
	Unresolved	<i>Inexact</i>	The resolved location is more generic than the annotation but is evaluated as useful.	<i>Myllypuro, a neighbourhood, instead of Myllypuro nursing home</i>	Geo-text data
		<i>Omission</i>	Toponym is missing from the gazetteer.	<i>Selkämeri (Bothnian sea)</i>	Noise/information loss
		<i>Exonym</i>	Toponym is an exonym, e.g., a Finnish variant of a toponym outside Finland. The exonym might be missing from the gazetteer, or the query does not prioritise alternative names.	<i>Käkisalmi (modern Priozersk, Russia), Kabulin lentokenttä (Kabul international airport).</i>	Noise/information loss

The possible outcomes (geo-text data, noise, information loss) of these errors are listed.



**Table 3.** Toponym recognition metrics sorted by *F1*-score on the two corpora with different toponym recognition models.

Toponym recognition model	Precision ↑	Recall ↑	<i>F1</i> -score ↑	Lematisation accuracy ↑
<i>News corpus</i>				
Fine-tuned XLM-RoBERTa	<b>72.6</b> (1.14)	77.89 (3.73)	<b>75.11</b> (1.91)	<b>85.71</b>
Fine-tuned FinBERT	71.55 (1.61)	<b>77.99</b> (3.40)	74.61 (2.19)	81.48
spaCy_large	68.68	66.14	67.39	76.19
spaCy_small	59.66	55.56	57.54	74.07
<i>Tweet corpus</i>				
Fine-tuned FinBERT	<b>84.27</b> (2.59)	<b>77.63</b> (3.81)	<b>80.71</b> (1.11)	83.94
Fine-tuned XLM-RoBERTa	81.96 (1.73)	65.42 (2.91)	72.73 (2.01)	<b>84.94</b>
spaCy_large	70.43	58.84	64.12	75.3
spaCy_small	53.72	47.79	50.58	69.88

Mean and (standard deviation) of five training runs are reported for the Transformer-based models. Lematisation accuracy, here, covers only toponyms and not all tokens. Arrows indicate whether a higher ↑ or lower ↓ score is better. Bold values indicate the best performing model on that metric.

**Table 4.** Toponym resolution results in an optimal case, in which all toponyms were correctly recognised and lemmatised.

Corpus	Ground-truth source	Median (km) ↓	Mean (km) ↓	AUC ↓	ACC@161 (%) ↑	ACC@BBOX (%) ↑	ACC@POLY (%) ↑	Resolved (%) ↑
News	GeoNames	73.64	545.97	0.4	73.6	93.26	92.7	96.3
	<i>WoF/OSM</i>	0	422.98	0.069	92.13			
Tweets	GeoNames	20.69	216.07	0.32	86.8	89.83	87.88	95.58
	<i>WoF/OSM</i>	0	161.31	0.099	92.86			

Error distance-based metrics have two rows indicating the source gazetteer for the ground-truth coordinates: GeoNames (top) or Who's on First/Open Street Map (bottom, *italicised*). Polygons and bounding boxes are always sourced from WoF/OSM. Arrows indicate whether a higher ↑ or lower ↓ score is better.

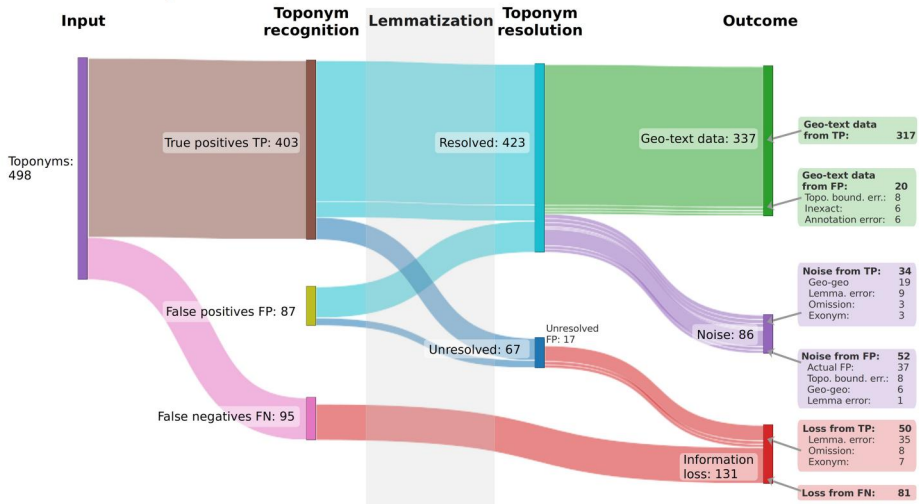
Toponym resolution results are reported in Table 4. The system resolves most locations correctly, as indicated by a high accuracy within the ground-truth polygons. Misaligned coordinates explain much of the poor resolution performance, especially on the news dataset. For example, ACC@161 performance improves by almost 20% on the news dataset when both the prediction and the ground-truth are from the same source, indicating a significant proportion of toponyms were misaligned by more than 161 km. Whether the ground-truth area is a full polygon, or a bounding box shows no significant difference on the news corpus. On the tweet corpus, the bounding boxes slightly overestimate performance. Roughly 5% of toponyms were left unresolved.

## 5.2. Error analysis results

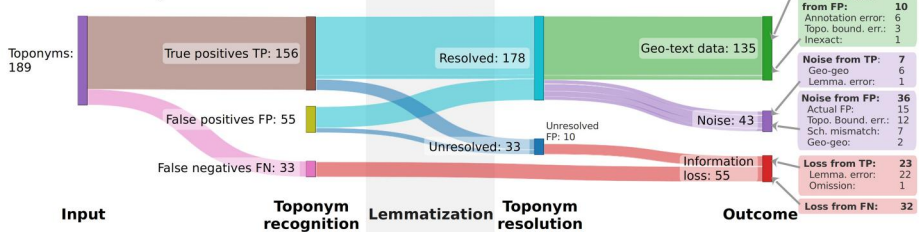
Figure 5 shows the full geoparsing process from input texts to output geo-text dataset. It visualises the outcome – geo-text data, noise or information loss – for each toponym. The majority of toponyms in both corpora are geoparsed successfully. However, major proportions of the output are noise: 26% for tweets and 32% for news. Similarly, 26% and 29% of possible geo-text data is lost. The geoparser functions slightly better on the tweet than the news dataset. The primary causes of the errors are similar on both corpora: the toponym recogniser returns noise, resolution fails due to geo-geo ambiguity, or the toponym is lemmatised incorrectly.



## A. Tweet corpus



## B. News corpus



**Figure 5.** A Sankey diagram illustrating the outcomes for each toponym for (A) tweet and (B) news corpora. The diagrams show whether each toponym is correctly recognised and resolved, and what are the causes of errors. See Table 2 for an explanation of the error causes.

## 6. Discussion

### 6.1. Generic tools simplified the development of a geoparser for Finnish

How well did the geoparser perform? The best Transformer-based models achieved toponym recognition *F1*-scores above 75 points on both corpora (Table 3). Roughly, 90% of toponyms were resolved within the correct area (Table 4). Although a like-for-like comparison is not possible, these results are similar to those achieved with generic resources on English corpora (Wang and Hu 2019a, Gritta *et al.* 2020, X. Hu *et al.* 2023a, 2023b). However, the metrics give only one perspective to geoparser performance.

A full error analysis and classification that we conceptualised and executed is beneficial in at least three ways (Figure 5). First, a user seeking to apply geo-text data (Y. Hu 2018) can better appraise whether the expected amounts of data loss and noise are acceptable for their use case. Second, it guides the development of the geoparser by highlighting where improvements are most needed. For example, a more robust lemmatiser would reduce the number of failed queries caused by incorrect lemmas. Finally, roughly, 20% of toponyms that were recognised as false positives, due to, e.g.

toponym boundary errors, produced valid data in the end. This highlights the value of examining and visualising the individual steps of the geoparsing pipeline, rather than focusing on evaluating toponym recognition or resolution separately with quantitative measures.

Using generic resources similarly to those we employed is not a novel idea, as named entity recognisers have been used in geoparsers (Karimzadeh *et al.* 2019) and performance comparisons (Wang and Hu 2019b, Gritta *et al.* 2020, X. Hu *et al.* 2023b). Similarly, disambiguation by population and generic entity linking software have been applied to toponym resolution (X. Hu *et al.* 2023a). Often these approaches are base-lines that predictably perform worse than the more intricate, customised solutions. Generic resources present a compromise – relying on them means accepting their limitations and sometimes ill-fitting design choices.

On the other hand, drawing from the wider NLP and GIS resources allowed us to focus on missing elements, such as annotating evaluation corpora. Significant investments into gathering training material, annotating evaluation corpora and implementing geoparsing solutions are needed for expanding geoparser coverage to a new language. Given the digital divide in linguistic resource availability (Bender 2019), resources may not exist at all for some target languages. Based on the satisfactory results achieved by our approach, we believe generic resources are a worthy compromise.

Our study has limitations. We have not presented comprehensive comparisons between alternative recognition and resolution methods. In addition, the evaluation datasets contain only about 700 toponyms in total, and previous comparisons show that performance varies between corpora (Wang and Hu 2019b). Although we limited our focus to Finnish, future research could test the validity of our approach for other languages or in a multilingual case.

## **6.2. Error distance metrics are insufficient for a nuanced understanding of toponym resolution performance**

We proposed measuring the proportion of geoparser predictions intersecting with ground-truth areas. Importantly for geoparsing corpora development, we found only minor differences between using polygons and bounding boxes derived from them. Areal metrics differ from common metrics in at least three significant ways. First, ground-truth coordinates from different gazetteers were misaligned by, on average, 146 km on the news corpus and 82 km on the tweet corpus. The effect of this misalignment is shown by the improvements across all metrics when the predicted and actual coordinates come from the same gazetteers (Table 4). ACC@POLY and ACC@BBOX are less sensitive to misalignment since an area should be roughly identical regardless of source. Second, toponyms are used to refer to locations of all scales, from local (points-of-interest, facilities) to global (mountain ranges, oceans, continents) (Table 1). Thus, the significance of error distance varies by toponym (Gritta *et al.* 2018a). A constant error distance cut-off value, such as 161 km is not sensitive to such variation and can therefore be both too lenient and overly strict (Figure 4, Table 4). ACC@161 overestimates system performance on the tweet corpus by about five points

compared to ACC@POLY (Table 4), whereas differences between the bounding box and polygon methods were less than 2 points. Finally, containment is more important than distance in certain cases. For example, the point representation of a city should fall within the city boundaries, even if points outside it yield lower error distances.

Thus, error distance may not be useful as a sole measurement and metrics derived from it may even mislead. Gritta *et al.* (2020) argued that multiple toponym resolution metrics are needed for a complete understanding of performance in different contexts. Given the importance of robust metrics for comparing geoparsers and defining the state-of-the-art (Wang and Hu 2019a, Gritta *et al.* 2020, X. Hu *et al.* 2023a), we believe toponym resolution evaluation needs alternatives that do not rely on unscaled error distances. ACC@POLY and ACC@BBOX are one such approach.

However, the geometric representation of both the predictions and ground-truth coordinates affect the usability of these metrics. ACC@POLY and ACC@BBOX assume the prediction is a point coordinate and the ground-truth has a polygon representation. Should the geoparser produce line or polygon geometry predictions, metrics based on areal overlap, such as those proposed by Laparra and Bethard (2020), are more suitable. If the ground-truth locations lack polygon representations – for example, they are point-of-interests or street segments – the error distances could be scaled as proposed by Gritta *et al.* (2018a).

### 6.3. Target language is relevant in geoparsing

The choice of target language is not trivial for the geoparsing task. For example, the error analysis of our system showed that lemmatisation and Finnish versions of place names absent from the gazetteer constitute major sources of error (Figure 5). Lemmatisation is less relevant when geoparsing languages with limited inflectional morphology, such as English. Therefore, there has been no reason to focus research efforts on it. In contrast, either further improving the lemmatiser or implementing robust gazetteer-free toponym resolution methods, which in turn may require a plethora of geo-text data as training material, is relevant to improve geoparsing for languages such as Finnish. In addition, the distribution of toponyms in supposedly global corpora show a marked emphasis on the English-speaking world (Figure S1) and resources for training and evaluation limit the development of geoparsers. Based on our results, we argue that target language is relevant at least in terms of research focus areas; text processing steps required; resource availability for the development and evaluation geoparsers; and the geographic distribution of locations used for evaluation.

Therefore, we argue that target language is too important to be assumed implicitly. We call for the adoption of the Bender rule in geoparsing: ‘Always name the language(s) you’re working on’ (Bender 2019). This simple proposition can offer a different angle on ongoing debates. For example, Wang and Hu (2019a) posed the question of whether geoparsers perform sufficiently well to consider geoparsing to be a ‘solved’ problem. Based on a systematic evaluation of geoparsers for English, they conclude that geoparsing is indeed solved for most common toponyms when the input texts are well-formatted (Wang and Hu 2019a). This conclusion was challenged

by Liu *et al.* (2022b) who evaluated toponym recognition and resolution spatially explicitly, and argue that geoparsing is not solved, but spatially biased. We argue that linguistic coverage is as important as spatially unbiased coverage. Thus, the question could be formulated ‘solved for which languages?’.

By emphasising the role of the target language in geoparsing, we do not intend to diminish advances made for any single language. Nor do we, by having focused our criticism towards English, wish to overlook geoparsing research on other languages, such as Chinese (Ma *et al.* 2022) and Indonesian (Dewandaru *et al.* 2020). Rather, we echo and emphasise the arguments that approaches may not be applicable cross-lingually (Bender 2011, 2019). Language cannot be ignored if geoparsing is to be effectively used across linguistic and geographic boundaries. In the future, more explorations into previously uncovered languages, multilingual geoparsing approaches (e.g. Pouliquen *et al.* 2004, Liu *et al.* 2022a), and means for producing novel evaluation corpora for them, e.g. programmatically (Gritta *et al.* 2018a) are needed.

## 7. Conclusions

In this article, we created a geoparser for Finnish texts with existing GIS and NLP resources. We annotated two corpora and evaluated the system with an extensive error analysis to test the validity of our approach. A systematic error analysis revealed failure points on all geoparsing processing steps. Besides common toponym recogniser and resolver failures, lemmatisation, which is trivial for English but relevant for our target language, caused a significant proportion of the errors. Evaluation also showed that our approach performs well on social media and news text domains, despite using generic resources. Drawing from the literature and our observations, we call for wider recognition for the role of target language in the geoparsing task. Noting the shortcomings of error distance metrics, we additionally proposed measuring the proportion of predictions intersecting ground-truth areas. We hope our findings help to highlight new aspects of geoparser development and work to emphasise geoparsers as tools for creating valid, thoroughly evaluated geo-text data.

## Notes

1. Rebranded as X in July 2023.
2. <https://www.openstreetmap.org/export>.
3. <https://whosonfirst.org/>.
4. <https://www.geonames.org/>.
5. <https://huggingface.co/Finnish-NLP/roberta-large-finnish-v2>.
6. [https://github.com/UniversalDependencies/UD\\_Finnish-TDT/tree/master](https://github.com/UniversalDependencies/UD_Finnish-TDT/tree/master).
7. <https://github.com/TurkuNLP/turku-one/tree/main>.
8. <https://spacy.io/usage/models>.

## Acknowledgements

The authors thank the three anonymous reviewers for their critical comments and Ross Purves for comments and discussions that greatly improved this article. The authors wish to acknowledge CSC – IT Center for Science, Finland, for providing computational resources. We made use

of geospatial tools acquired via the Geoportti RI (Open Geospatial Information Infrastructure for Research, urn:nbn:fi:research-infras-2016072513).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by Kone Foundation (project MOBICON) and the Emil Aaltonen Foundation.

## Notes on contributors

**Tatu Leppämäki** is a doctoral researcher in the Digital Geography Lab at the Department of Geosciences and Geography, University of Helsinki, Finland. His research is focused on method development for structuring big data with a focus on geographical and linguistic data. He contributed to conceptualisation, methodology, investigation, software development, data curation, visualisation and writing.

**Tuuli Toivonen** is a Professor in Geoinformatics and leads the Digital Geography Lab at the Department of Geosciences and Geography, University of Helsinki, Finland. Her research explores the opportunities provided by using mobile big data and other novel data sources to support spatial planning and decision-making towards fair and sustainable societies. She contributed to conceptualisation, writing (review and editing), supervision, project administration and funding acquisition.

**Tuomo Hiippala** is a Professor of English Language and Digital Humanities at the University of Helsinki, Finland. His current research interests include computational analysis of multimodal communication and applications of language technology in the humanities and social sciences. He contributed to conceptualisation, methodology, supervision and writing (review and editing).

## ORCID

Tatu Leppämäki  <http://orcid.org/0000-0002-9634-7943>

Tuuli Toivonen  <http://orcid.org/0000-0002-6625-4922>

Tuomo Hiippala  <http://orcid.org/0000-0002-8504-9422>

## Data and codes availability statement

An up-to-date version of the geoparser used in this work is available at the link: <https://github.com/DigitalGeographyLab/Finger-geoparser>

The evaluation corpora, data, code and instructions for replicating the findings of this study are available at the link: <https://doi.org/10.6084/m9.figshare.25968280.v1>

## References

- Acheson, E. and Purves, R.S., 2021. Extracting and modeling geographic information from scientific articles. *PLOS One*, 16 (1), e0244918.
- Acheson, E., Sabbata, S.D., and 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.

- Amitay, E., et al., 2004. Web-a-where: geotagging web content. In: *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*, 25–29 July 2004 Sheffield, United Kingdom. New York: Association for Computing Machinery, 273–280.
- Ardanuy, M.C., et al., 2023. The past is a foreign place: improving toponym linking for historical newspapers. In: A. Sela, F. Jannidis, and I. Romanowska, eds. *Proceedings of the computational humanities research conference 2023*, 6–8 December 2023, Paris, France, 368–390. Available from: <https://ceur-ws.org/Vol-3558/paper4426.pdf>
- Avvenuti, M., et al., 2018. CrisMap: a big data crisis mapping system based on damage detection and geoparsing. *Information Systems Frontiers*, 20 (5), 993–1011.
- Bender, E., 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6 (3), 1–26.
- Bender, E., 2019. The #BenderRule: on naming the languages we study and why it matters. *The Gradient*. Available from: <https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/>
- Berragan, C., et al., 2022. Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science*, 37 (4), 747–766.
- Blasi, D.E., et al., 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26 (12), 1153–1170.
- Conneau, A., et al., 2020. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5–10 July 2020 Online. Stroudsburg: Association for Computational Linguistics, 8440–8451.
- Devlin, J., et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North*, 2–7 June 2019 Minneapolis, United States. Stroudsburg: Association for Computational Linguistics, 4171–4186.
- Dewandaru, A., Widyantoro, D.H., and Akbar, S., 2020. Event geoparser with pseudo-location entity identification and numerical argument extraction implementation and evaluation in Indonesian news domain. *ISPRS International Journal of Geo-Information*, 9 (12), 712.
- Fize, J., Moncla, L., and Martins, B., 2021. Deep learning for toponym resolution: geocoding based on pairs of toponyms. *ISPRS International Journal of Geo-Information*, 10 (12), 818.
- Gelernter, J. and Mushegian, N., 2011. Geo-parsing messages from microtext. *Transactions in GIS*, 15 (6), 753–773.
- Gregory, I., et al., 2015. Geoparsing, GIS, and textual analysis: current developments in spatial humanities research. *International Journal of Humanities and Arts Computing*, 9 (1), 1–14.
- Gritta, M., et al., 2018a. What's missing in geographical parsing? *Language Resources and Evaluation*, 52 (2), 603–623.
- Gritta, M., Pilehvar, M.T., and Collier, N., 2018b. Which Melbourne? Augmenting geocoding with maps. In: *Proceedings of the 56th annual meeting of the association for computational linguistics*, 15–20 July Melbourne, Australia. Stroudsburg: Association for Computational Linguistics, 1285–1296.
- Gritta, M., Pilehvar, M.T., and Collier, N., 2020. A pragmatic guide to geoparsing evaluation: toponyms, named entity recognition and pragmatics. *Language Resources and Evaluation*, 54 (3), 683–712.
- Hahmann, S. and Burghardt, D., 2013. How much information is geospatially referenced? Networks and cognition. *International Journal of Geographical Information Science*, 27 (6), 1171–1189.
- Halterman, A., 2023. *Mordecai 3: a neural geoparser and event geocoder*. arXiv:2303.13675. <https://arxiv.org/abs/2303.13675>
- Hu, X., et al., 2023a. How can voting mechanisms improve the robustness and generalizability of toponym disambiguation? *International Journal of Applied Earth Observation and Geoinformation*, 117, 103191.
- Hu, X., et al., 2023b. Location reference recognition from texts: a survey and comparison. *ACM Computing Surveys*, 56 (5), 1–37.

- Hu, Y. and Adams, B., 2021. Harvesting big geospatial data from natural language texts. In: M. Werner, and Y.Y. Chiang, eds. *Handbook of big geospatial data*. 1st ed. Cham: Springer, 487–508.
- Hu, Y., 2018. Geo-text data and data-driven geospatial semantics. *Geography Compass*, 12 (11), e12404.
- Hulden, M., Silfverberg, M., and Francom, J., 2015. Kernel density estimation for text-based geolocation. In: *Proceedings of the twenty-ninth AAAI conference on artificial intelligence*, 25–30 January 2015 Austin, United States. Palo Alto: AAAI Press, 145–150.
- Janowicz, K., et al., 2022. Six GIScience ideas that must die. *AGILE: GIScience Series*, 3, 1–8.
- Karimzadeh, M., et al., 2019. GeoTxt: a scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23 (1), 118–136.
- Kulkarni, S., et al., 2021. Multi-level gazetteer-free geocoding. In: *Proceedings of second international combined workshop on spatial language understanding and grounded communication for robotics*, 5–6 August 2021 Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics, 79–88.
- Laparra, E. and Bethard, S., 2020. A dataset and evaluation framework for complex geographical description parsing. In: *Proceedings of the 28th international conference on computational linguistics*, 8–13 December 2020 Barcelona, Spain. International Committee on Computational Linguistics, 936–948.
- Leidner, J. and Lieberman, M., 2011. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3 (2), 5–11.
- Leppämäki, T., 2022. *Developing a Finnish geoparser for extracting location information from unstructured texts*. Unpublished MA thesis. University of Helsinki.
- Lieberman, M.D., Samet, H., and Sankaranarayanan, J., 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In: *2010 IEEE 26th international conference on data engineering (ICDE 2010)*, 1–6 March 2010 Long Beach, United States. Los Alamitos: IEEE Computer Society, 201–212.
- Liu, P., et al., 2022a. Extracting locations from sport and exercise-related social media messages using a neural network-based bilingual toponym recognition model. *Journal of Spatial Information Science*, 24 (24), 31–61.
- Liu, Z., et al., 2022b. Geoparsing: solved or biased? An evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3, 1–13.
- Luoma, J., et al., 2021. Fine-grained named entity annotation for Finnish. *Proceedings of the 23rd Nordic conference on computational linguistics (NoDaLiDa)*, 135–144. Available from: <https://aclanthology.org/2021.nodalida-main.14>
- Ma, K., et al., 2022. Chinese toponym recognition with variant neural structures from social media messages based on BERT methods. *Journal of Geographical Systems*, 24 (2), 143–169.
- Middleton, S.E., et al., 2018. Location extraction from social media: geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, 36 (4), 1–27.
- Moncla, L., et al., 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In: *Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems*, 4–7 November 2014 Dallas, United States. New York: Association for Computing Machinery, 183–192.
- Montani, I., et al., 2023. *Explosion/spaCy: V3.5.1: Spacat for multi-class labeling, fixes for textcat + transformers and more (v3.5.1)* [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.7715077>
- Poulliquen, B., et al., 2004. Geographical information recognition and visualization in texts written in various languages. In: *Proceedings of the 2004 ACM symposium on applied computing – SAC '04*, 14–17 March 2004 Nicosia, Cyprus. New York: Association for Computing Machinery, 1051.
- Purves, R.S., et al., 2018. Geographic information retrieval: progress and challenges in spatial search of text. *Foundations and Trends® in Information Retrieval*, 12 (2–3), 164–318.
- Purves, R.S., Winter, S., and Kuhn, W., 2019. Places in information science. *Journal of the Association for Information Science and Technology*, 70 (11), 1173–1182.



- Pyysalo, S., et al., 2015. Universal dependencies for Finnish. In: *Proceedings of the 20th Nordic conference of computational linguistics (Nodalida 2015)*, 11–13 May 2015 Vilnius, Lithuania. Sweden: Linköping University Electronic Press, 163–172.
- Tkachenko, M., et al., 2020. *Label Studio: data labeling software*. Available from: <https://github.com/heartexlabs/label-studio>
- Tobin, R., et al., 2010. Evaluation of Georeferencing. In: *Proceedings of the 6th workshop on geographic information retrieval (GIR'10)*, 18–19 February 2010 Zurich, Switzerland. New York: Association for Computing Machinery, 7.
- Virtanen, A., et al., 2019. *Multilingual is not enough: BERT for Finnish*. arXiv Preprint arXiv: 1912.07076.
- Wallgrün, J.O., et al., 2018. GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32 (1), 1–29.
- Wang, J. and Hu, Y., 2019a. Are we there yet? Evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. In: *Proceedings of the 3rd ACM SIGSPATIAL international workshop on geospatial humanities*, 5 November 2019 Chicago, United States. New York: Association for Computing Machinery, 1–6.
- Wang, J. and 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.
- Wang, J., Hu, Y., and Joseph, K., 2020. NeuroTPR: a neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS*, 24 (3), 719–735.
- Weischedel, R., et al., 2013. OntoNotes release 5.0 (p. 2806280 KB) [dataset]. Linguistic Data Consortium. <https://doi.org/10.35111/xmhb-2b84>