medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Mapping the plague through natural language processing

Fabienne Krauer[1,2], Boris V. Schmid[1]

[1]*Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, NO-0316 Oslo, Norway*

[2] *Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, UK*

Abstract

Plague has caused three major pandemics with millions of casualties in the past centuries. There is a substantial amount of historical and modern primary and secondary literature about the spatial and temporal extent of epidemics, circumstances of transmission or symptoms and treatments. Many quantitative analyses rely on structured data, but the extraction of specific information such as the time and place of outbreaks is a tedious process. Machine learning algorithms for natural language processing (NLP) can potentially facilitate the establishment of datasets, but their use in plague research has not been explored much yet. We investigated the performance of five pre-trained NLP libraries (Google NLP, Stanford CoreNLP, spaCy, germaNER and Geoparser.io) for the extraction of location data from a German plague treatise published in 1908 compared to the gold standard of manual annotation. Of all tested algorithms, we found that Stanford CoreNLP had the best overall performance but spaCy showed the highest sensitivity. Moreover, we demonstrate how word associations can be extracted and displayed with simple text mining techniques in order to gain a quick insight into salient topics. Finally, we compared our newly digitised plague dataset to a re-digitised version of the famous Biraben plague list and update the spatio-temporal extent of the second pandemic plague mentions. We conclude that all NLP tools have their limitations, but they are potentially useful to accelerate the collection of data and the generation of a global plague outbreak database.

Introduction

The dissemination of historical plague across the globe is being studied and discussed since many decades. The second pandemic entered Europe in 1347 [1], but it may have started a century earlier in Central Asia and China [2]. The pandemic caused millions of casualties before gradually disappearing from Europe and the Mediterranean in the 18th century. The third pandemic is thought to have started towards the end of the 18th century in Yunnan (China) [3]. It reached Hong Kong in 1894 from where it spread globally. Since the 19th century - and perhaps even earlier - many scholars have collected data on plague outbreaks in a more or less systematic manner. Among the earliest compilations with a broader geographical coverage are the works of Hecker in 1865 [4], Martin in 1879 [5], Creighton in 1891 for the British Isles [6], Dörbeck in 1906 for Russia [7] or Sticker in 1908 for the whole world [8]. In 1975, Biraben combined data from some of these and many other sources into the largest tabellaric collection of yearly plague outbreaks published to date [9]. His list has been used by many for spatio-temporal analyses of plague dispersion in Europe [10-14]. In the decades after Biraben, many more scholars have continued his legacy and collected, reviewed and discussed historical data for selected regions or periods, for example Marien [15], Varlik [16] and

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Panzac [17] for the Ottoman Empire, Noordegraaf [18] and Roosen [19] for the Low Countries, Dols for the Black Death in the Middle East [20, 21], Slavin for the *pestis secundis* [22], Frandsen for the 1709-13 plague in the Baltics [23], Panzac for the islands in the eastern Mediterranean in the 17th to 19th century [24] or Fazlinejad and Ahmadi for 14th and 15th century Iran [25]. Most of these publications (both recent or more historical) contain information about the places and times of epidemics, which are the core aspects of infectious disease epidemiology. Many offer also additional information such as the size of the outbreaks, symptoms, treatments, control measures, putative transmission routes and ecological aspects. The body of publications is large, but narrative text must be converted into quantitative data in order to be usable for epidemiological analyses. However, the extraction of data from running text is time and labour intensive.

In the past few years, advances in machine learning algorithms and increasing computing efficiency have led to a rise of digital methods in epidemiology [26]. Particularly the automated generation of data from text through Natural Language Processing (NLP) has gained popularity. NLP approaches have been used to collect clinical information from patient health records (reviewed in [27]) or to analyze the spread of infectious diseases based on social media postings [28]. NLP algorithms have also been applied on various historical documents and text corpora (for an overview see [29]), for example to investigate the relationship between trade commodities and geographical locations [30] or to analyze the geographical distribution of cholera mentions in the UK Registrar General's reports from England and Wales in the 19th century [31]. The plague dot txt project at the University of Edinburgh has recently started to develop a NLP workflow to build a structured account of plague epidemiology based on treatises and publications about the third pandemic [32]. To our knowledge, the latter is the only project to date that explores the use of NLP in plague research.

The possibilities of NLP algorithms are manifold. They can partition a text word-wise (tokenization), attribute inflected forms of tokens to their canonical form (lemmatization), analyze the syntax (position-of-speech, POS), identify entities (named entity recognition, NER) or analyze the sentiment. The POS analysis partitions a running text into tokens (usually singular words) and returns information about the morphological class of each token (e.g. nouns, verbs). The NER analysis identifies and classifies tokens or combinations of tokens into pre-defined categories based on rules (i.e. a dictionary), statistical predictions, or both. A special case of NLP NER is the extraction of geographical data from a text (geoparsing). Geoparsing consists of two main steps: 1. Tagging, i.e. identification of a geographical entity (toponym), and 2. Geocoding, i.e. linkage of the geographical entity with GIS data such as coordinates. The GIS information is usually looked up in a geographical gazetteer. In theory, both steps can be done by hand and/or separately, but automated workflows may be preferable because they are faster and potentially more reproducible. General NLP libraries have to be combined with a geocoding service to deliver the same results as a designated geoparser.

In general, text mining tools can accelerate the generation of large datasets, but their performance has to be sufficient to outweigh the errors arising from the automated process. The performance of these algorithms depends on the chosen model or algorithm, and the structure and language of the text. Ideally, an NLP algorithm has a high recall or sensitivity (e. g. the proportion of locations that are correctly identified as locations) and a high specificity (e. g. the proportion of non-locations that are correctly identified as non-locations). Various NLP algorithms and libraries have been tested for modern English medical and non-medical texts and their performances differ substantially (see e.g. [27, 33]). The literature on performance evaluation of NLP libraries for more historical texts is sparser.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

For example the sensitivity and the precision of the Edinburgh Geoparser, a popular tool for historical English texts, was found to vary between 60 and 80% depending on the text [34].

There is a growing scientific interest in building a global database of historical plague outbreak [35]. The Black Death Digital Archives project (http://globalmiddleages.org/project/black-death-digital-archive-project) initiated by Green and Roosen aims to "newly interrogate our traditional sources of historical information" and to link biological, archaeological and documentary databases [36]. We here contribute to this effort with a case study on the use of NLP to facilitate the digitisation of plague location data. We use a German plague treatise by Sticker [8] as an example. Sticker's work was one of the main sources Biraben used for his compilation, and it potentially contains additional information that can expand our knowledge of the historical, spatio-temporal spread of plague. We investigate and compare the application of different NLP libraries for the extraction of places with plague and explore how to illustrate often discussed topics in the book with simple word embedding and network plotting. Finally, we present a novel, geocoded plague dataset based on Sticker's book and compare it to a re-digitised version of Biraben's plague data set.

Methods

Source text

Our source text is a German plague treatise ("Geschichte der Pest") published in 1908 by Georg Sticker [8]. Sticker was a German physician, who together with Robert Koch was sent to Bombay to investigate the plague epidemic in 1897. In the first part of the book, he narrates the historical spread of plague on five continents (Europe, Asia, the Americas, Africa and Australia) in chronological order from biblical times to 1908. In the second part, Sticker elaborates on plague ecology, microbiology and pathology. For the historical part, he combined information from secondary literature, but he also consulted original sources (see his introduction). It is divided in 16 Chapters, of which the first four chapters are about the first pandemic, chapters 5 to 15 are about the second pandemic and the chapters 15 and 16 refer to the third pandemic. For our analysis, we focus on plague during the second and third pandemic, corresponding to the chapters five to sixteen of Sticker's treatise on plague (pages 42 to 399). The structure of the text is a combination of running text interspersed with semi-tabulated year and place listings. The running text contains both specific information about places that mention plague in a given year, but also general information on plague as well as historical anecdotes and elaborations. A scanned OCR version of the book is freely available on the Internet Archive (https://archive.org/details/ abhandlungenausd01stic/mode/2up).

Preprocessing and establishment of gold standard

In a first preprocessing step, we cleaned the raw OCR text manually. We removed interspersed tables, page numbers and page headers, and corrected misaligned text. We also removed end-of-line hyphenations and notes in the book margins that were erroneously included in the running text. We checked the text file for OCR errors by looking for special characters and words that were not recognized by the Notepad++ Spell Checker. We then established the gold standard dataset of location toponyms, with both authors independently annotating the preprocessed text using the annotator tool webanno (version 3.5.9) [37]. We then compared the two annotations and established a consensus document. This list of toponyms contained all geographical entities in the text irrespective of whether the location was linked to plague or not. We included all administrative place, region or country

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

names as well as natural features such as "the Black Sea". Associative toponyms such as "the Bishop of Avignon" were excluded because they are not true locations. This gold standard list was used for the evaluation of the tagging performance of various NLP libraries (see below). We then used this list to generate the final dataset of places with plague outbreaks. For this we extracted text snippets of 50 characters before and after each toponym to obtain the context and decided for each case individually whether it was linked to a specific plague outbreak. Furthermore, we also extracted the corresponding years (usually a four-digit string) using regular expression (regex) and allocated them manually to the corresponding toponym. We also linked the referenced author names (i.e. the source of the information) with the corresponding places wherever it was available. Finally, we batch geocoded these locations using the REST API services of both ArcGIS (https://developers.arcgis.com/rest/) and Google Geocoding (https://developers.google.com/maps/documentation/geocoding/start) to query the GIS information for each place. We extracted the modern place names, the country ISO code, the centroid and bounding box coordinates and the type of administrative unit. The bounding box coordinates are the minimum and maximum longitudes and latitudes of a given administrative unit, and can be used a proxy for the spatial extent of a place. All coordinates are provided in WGS84. We then compared both sets of results and chose the better match in terms of centroid and bounding box coordinates as the final result in our dataset. Ambiguous or unclear toponyms or questionable results were checked individually by consulting the original literature or other sources referenced therein. Historical or colloquial regions without a clear administrative border were geocoded approximatively by defining the boundary coordinates manually based on maps on Wikipedia and calculating the arithmetic centroid coordinates. Toponyms that could not be localized exactly were geocoded according to the next lower identifiable level administrative unit and were marked as approximate. Toponyms that could not be localized at all were marked as unknown. The definition of the administrative units returned by ArcGIS and Google varied by country, and we thus re-categorized all results as one of the following: place (city, town, village, neighborhood, district, municipality and other populated place), administrative unit (county, state and province), country, island and region (colloquial area, historical or geographical region, and natural features such as streams, mountains or lakes). This gold standard dataset was used for the performance evaluation of the geocoding algorithms and is also the final output of our study. The study was conducted in a Windows environment with a german locale. All work was carried out in R/R Studio (version 4.0.0) and Notepad++. The R code and the final plague datasets are available in a repository (https://doi.org/10.5281/zenodo.4724016) [38].

Toponym tagging performance evaluation

We tested four different NLP libraries and one geoparser for the tagging of toponyms: Google NLP [39], Stanford CoreNLP [40] with the pre-trained German model version 2018-10-05 [41], spaCy [42] with the pre-trained German model version 2.1.0 [43], germaNER [44] and Geoparser.io [45]. For a technical comparison of the libraries see supplement Table S2. We performed syntax analysis (POS) and entity recognition (NER). Google NLP differentiates between person, location, organization, event, work of art, consumer good and other, while Stanford CoreNLP differentiates between person, location, organization, misc and "o" (outside, not classified). The german model used with SpaCy as well as germaNER both differentiate between location, person, organization and misc/other. Geoparser.io only returns toponyms and the corresponding GIS information but not the tokenization of the complete text. All algorithms accept running text except germaNER, which requires a priori tokenization. We therefore used the tokenization returned by spaCy as an input for germaNER. The

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

german Stanford CoreNLP java library (version 2018-10-05) was downloaded from the Stanford NLP Github Page (https://stanfordnlp.github.io/CoreNLP/human-languages.html) and accessed through the R package coreNLP (version 0.4.2) [46]. SpaCy (v2.0) was downloaded and accessed through the R package spacyr (version 1.2) [47]. The java standalone for GermaNER was downloaded from the Github account (https://github.com/tudarmstadt-lt/GermaNER) and run from the command line. To facilitate the automated geoparsing approach, we removed all words or sentences in parentheses, which were mainly author names and references and thus irrelevant for the tagging.

We then assessed the performance to identify toponyms of each of the five approaches compared to the gold standard using various indicators. For this, we first combined all results and the gold standard in one dataset by mapping all the entities onto the tokens returned by spacy. This was necessary because the original tokenization differed between the different libraries making the comparison difficult. Spacy was chosen as the base because it has the most stringent tokenization pattern. After the mapping, we re-categorized the entities of all five approaches as “location” or “other” (which includes not identified tokens). If geographical entities were not recognized completely by a text mining algorithm, we allowed also for partial matches for the calculation of the performance. For example, “Freiburg im Breisgau” could be identified as “Freiburg” or the full name. We established for each token and for each approach whether it was a true positive (TP), a true negative (TN), a false positive (FP, type I error) or a false negative (FN, type II error). From this information we calculated several performance indicators, all of which range between 0 (poorest performance) and 1 (perfect performance). The accuracy is the overall proportion of correct predictions (positive and negative) among all predictions. The sensitivity (recall) is a measure of how good the algorithm is at correctly detecting locations. The specificity (selectivity) estimates the probability of correctly detecting a non-location. The positive predictive value (PPV or precision) is the proportion of predicted locations that are true locations. The negative predictive value (NPV) does the inverse, i.e. it estimates the probability that if a token is identified as a non-location, it is a true non-location. Of note, PPV and NPV depend on the proportion of locations in the text, i.e. they are intrinsic and cannot be compared between different texts. The F1 score is a harmonic average of sensitivity and precision with 0 being the lowest performance and 1 being the perfect performance. Finally, Cohen’s Kappa coefficient measures the observed accuracy compared to the expected accuracy when all agreement is by chance (random). A coefficient of 1 indicates perfect agreement; a coefficient of 0 indicates no agreement. The formal definition of all measures is given in supplement Table S1.

Geocoding performance evaluation

We also assessed the performance of two alternative geocoding services: Geoparser.io [45], which combines the tagging and geocoding, and Geonames [48]. Geoparser.io returns only the name of the toponym, the type and the centroid coordinates. Geonames.io provides more GIS information such as lower level administrative area units and place names in local or alternative languages. Geoparser.io returns the best match (according to internal criteria), while Geonames returns all possible matches in a ranked order. To make the algorithms comparable we picked only the first (i.e. best) match returned by Geonames. However, we restricted the Geonames search to places (P), administrative units (A), areas (L) and natural features (T, H and V). If no full match was found, we accepted also partial (fuzzy) match for Geonames. We then compared the performance of these two services to the geocoded gold standard dataset. For this, we combined the three datasets and calculated the Euclidean distances between the three centroid coordinates for each toponym. We assessed the

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

performance only for exactly located entities. We considered two places a match if both types were a country and the country ISO codes agreed. For entities that were not countries we considered it a match if the standard and comparator were in the same country and the Euclidian distance between the centroids of the standard and comparator was less than 30 km (for small entities with a standard bounding box up to 30 km), or less than half of the bounding box diameter of the standard (for larger entities with a standard bounding box diameter of more than 30 km). Based on the count of matches we calculated the proportion of toponyms identified (i.e. whether there was a result nor not) and the proportion of toponyms correctly identified for each approach. We also examined the mismatches and checked whether there was a potential regional or other bias in the geocoding. All Geocoding services were accessed through their REST-APIs between September and October 2019 using a designated batch geocoding script.

Exploratory text mining

Plague treatises commonly contain a wealth of information beyond the simple places and times of plague outbreaks, and we can gain additional insight by investigating salient topics and contexts. Our main interest was the occurrence of the most frequent nouns throughout the book (and thus over time), and their connection to other nouns. The declension of german nouns results in multiple word endings for the same word, and we thus used the lemmata instead of the tokens. After improving the lemmatization with a german morphological dictionary [49], we calculated the frequencies of all lemmata and chose the nine most frequent lemmata related to symptoms, infrastructure or disease transmission for further analysis as well as the word for mouse, which was historically often used as a synonym for rats. We then aimed to learn more about the context of these most frequent lemmata. Context cannot be extracted straightforwardly from a text, but we can use word embeddings (i.e. neighbor words) to investigate which words occur together frequently. For this, we constructed a co-occurrence matrix, which is a technique that counts how often any two words occur together within a given window. The matrix was constructed with a window of five words before and after each lemma. We then chose the ten most frequently co-occuring words for each of the selected nouns above and visualized the connection between them with a network plot.

Data description

Finally, we summarized the spatial and temporal coverage of our data set and compared it with a re-digitised version of Biraben's list (see supplemental Text S1). For this, we merge the two datasets by year and centroid coordinates. We calculated the proportion of full matches among all observations of both datasets for the same time period, plotted all locations in both datasets and compared the corresponding time series. We then restricted the merged dataset to the time period of the second pandemic and to exactly localized places (without regions, countries or other administrative areas) to update and summarise the spatio-temporal extent of the plague mentions.

Results

Gold standard

The OCR-corrected text of chapters five to sixteen of Sticker's treatise on plague was 864,106 characters long. Removing the author citations that are present throughout the text reduced the length

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

to 842,918 characters. We identified 7884 geographical entities with manual annotation (i.e. the gold standard). Of these 7884 toponyms only 4474 (57%) referred to a specific plague outbreak in a specific year (Figure 1). The rest were mainly repeated mentions of the same locations for a given year or additional geographical information to describe a place (e.g. "Geverske *near* Ostrovizza *in the region of* Zara"). Of these 4474 toponyms, 4087 (91.4%) could be localized exactly. Eight toponyms (0.2%) could not be localized at all ("unknown"). The remaining 379 toponyms (8.5%) were either colloquial or historical regions without clearly defined modern boundaries (271, e.g. "Podolia"), or populated places that could not be localized exactly but were attributed to a lower level administrative unit. These are marked as "approximate" in the final dataset.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

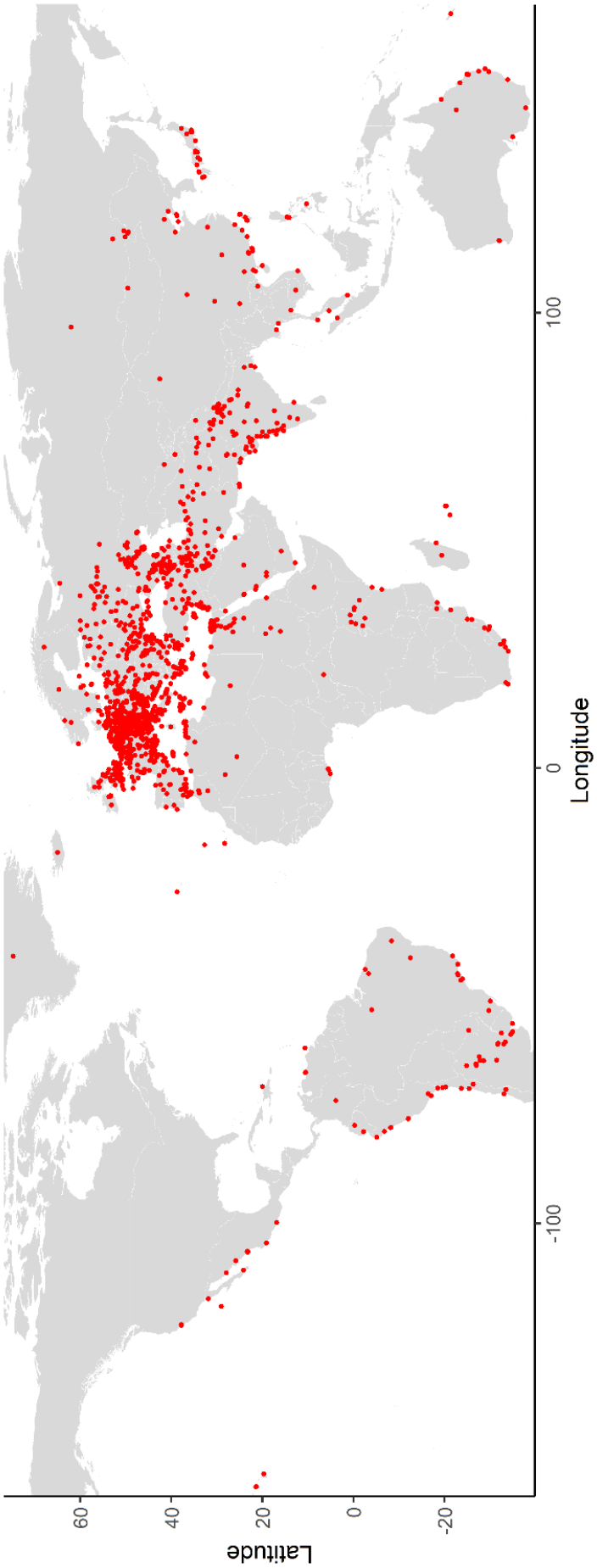

Figure 1. Spatial coverage of all geocoded geographical units (exact and approximate). Note that the data include different administrative levels from village to country, and the dots denote the centroids of each geographical entity.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Toponym tagging performance evaluation

The spaCy and the Stanford CoreNLP tokenizers yielded a similar number of total tokens (146,766 and 146,743 respectively) while Google NLP returned less tokens (146,340) (Table S3). GermaNER identified the most entities (34% of all recognized tokens = 50,374), followed by Google NLP (23% of all recognized tokens = 33,925), spaCy (9% of all recognized tokens = 12,963), Stanford coreNLP (6.5% of all recognized tokens = 9522) and Geoparser.io (3563). We then used the tokenization returned by the spaCy algorithm as the base and mapped all results of the NER analysis onto the tokens to compare the different approaches. With the gold standard, 5.7% of all tokens were classified as locations. The text mining algorithms of Google, Stanford, spaCy and germaNER resulted in 6.5%, 4.8%, 6.9% and 4.0% of the tokens categorized as locations. The Geoparser.io algorithm mapped only 2.5% of the tokens as locations.

We then evaluated the performance of the five text mining algorithms compared to the Gold standard for the detection of locations (Table 1). Overall, the proportion of correctly identified entities (accuracy) was large for all five libraries (range 0.96-0.99). The spaCy library showed the best sensitivity (0.92), followed by Stanford CoreNLP (0.81), Google (0.78), germanNER (0.61) and Geoparser.io (0.41). The specificity was equally high for all algorithms (range 0.98-0.99). Stanford coreNLP had the highest precision (PPV, 0.95, i.e. 5% false positives.), followed by Geoparser.io (0.91), germaNER (0.86), spaCy (0.76) and Google (0.67). The F1 scores and Cohen's kappa coefficients suggested a good overall performance for Stanford CoreNLP (0.87 and 0.87) and spaCy (0.83 and 0.82), a mediocre overall performance for Google NLP (0.72 and 0.70) and GermanER (0.71 and 0.70) and a poor performance for Geoparser.io (0.56 and 0.55). Of note, the slightly higher false positive rate of Google NLP can be explained by its rather broad definition of "location", which includes not only geographical entities but also nouns related to physical locations such as "Stadt" (town, city), "Provinz" (province) or "Haus" (house). Of the 8313 location tokens, only 26% were correctly identified by all algorithms. Locations that were missed by all algorithms included Germanized spelling (e.g. "Hoschiarpur" for "Hoshiarpur"), latin spelling (e.g. "Centumcellae" for "Civitavecchia"), composite entities (e.g. "Gurjewscher Kreis"), historic regions (e.g. "Podolien") or ambiguous words (e.g. "Sind" is a location but also a conjugated verb form of "to be"). The only token that was falsely identified as a location by all algorithms was "Santa Maria", which can be both a church name as well as a place name.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 1. Performance of different NLP algorithms for the identification of toponyms (location nouns).

	Google NLP	Stanford CoreNLP	spaCy	germaNER	Geoparser
TP	6451	6708	7642	5056	3381
FP	3154	329	2445	855	340
TN	135,229	138,124	136,008	137,598	138,113
FN	1862	1605	671	3257	4932
Accuracy	0.97	0.99	0.98	0.97	0.96
Sensitivity	0.78	0.81	0.92	0.61	0.41
Specificity	0.98	0.99	0.98	0.99	0.99
PPV	0.67	0.95	0.76	0.86	0.91
NPV	0.99	0.99	0.99	0.98	0.97
F1 score	0.72	0.87	0.83	0.71	0.56
Cohen's Kappa	0.70	0.87	0.82	0.70	0.55

Geocoding performance evaluation

To evaluate the geocoding performances, we compared the 4087 locations from the gold standard that were true plague locations and that were identifiable with the exact coordinates. Geonames identified substantially more toponyms (86.6 %) than Geoparser.io (50.1%). However, both geocoding services performed similarly well in the correct identification of a location (Geoparser.io 85.2% and Geonames 82.2%). Of note, the performance could be improved substantially if additional information was provided such as the country or region. However, for this study we evaluated the situation where there is no additional information available and where the performance depends on the best guess of the Geocoding algorithm. Many of the mismatches occurred for regions where places were renamed as the ruling power changed, through colonization or the contraction and expansion of empires.

Exploratory text mining

As shown in Figure 2A, outbreaks in towns were discussed more than twice as much as outbreaks in villages. Surprisingly, the word for clothes occurred as often as the word for rats. Figure 2B shows that the dissemination by ships was featured mostly in the 14th-15th and from the mid-17th to 19th century. Quarantine occurred from 1500 and onwards. The discussion of rats and mice was generally sparse until the end of the 19th century when the third pandemic started. It was at this time when the relationship of rats, fleas and plague transmission was discovered by Simond [50]. It is however unclear, whether in earlier epidemics dying rodents were rarely mentioned because the connection to plague transmission was unknown at the time or because they were not the main actors in the plague transmission cycle. We also explored how these most frequent topics related to each other and other lemmata with a word embedding approach. The resulting word associations are shown in Figure 2C. As anticipated, the words plague, village, town, human and year co-occurred often. We also found a word cluster describing the symptoms of plague (buboes, carbuncle, petechiae, fever). The "evil" was also often associated with the arrival of ships and sailors in the ports, which often triggered quarantine. Interestingly, the word for clothes was strongly connected to house, sick people, bed, town, equipment

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

and contagion. Clothes as potential sources of transmission are often mentioned in various historical sources. A famous example is the 1665 epidemic in Eyam, where the arrival of a bale of cloth from London supposedly caused the outbreak [51]. Alternative transmission routes by human ectoparasites have been repeatedly hypothesized for the second pandemic [52]. Given that human fleas and body lice are often found in clothes and bedding, the association of clothes with plague transmission may provide additional evidence in favour of the human ectoparasites hypothesis.

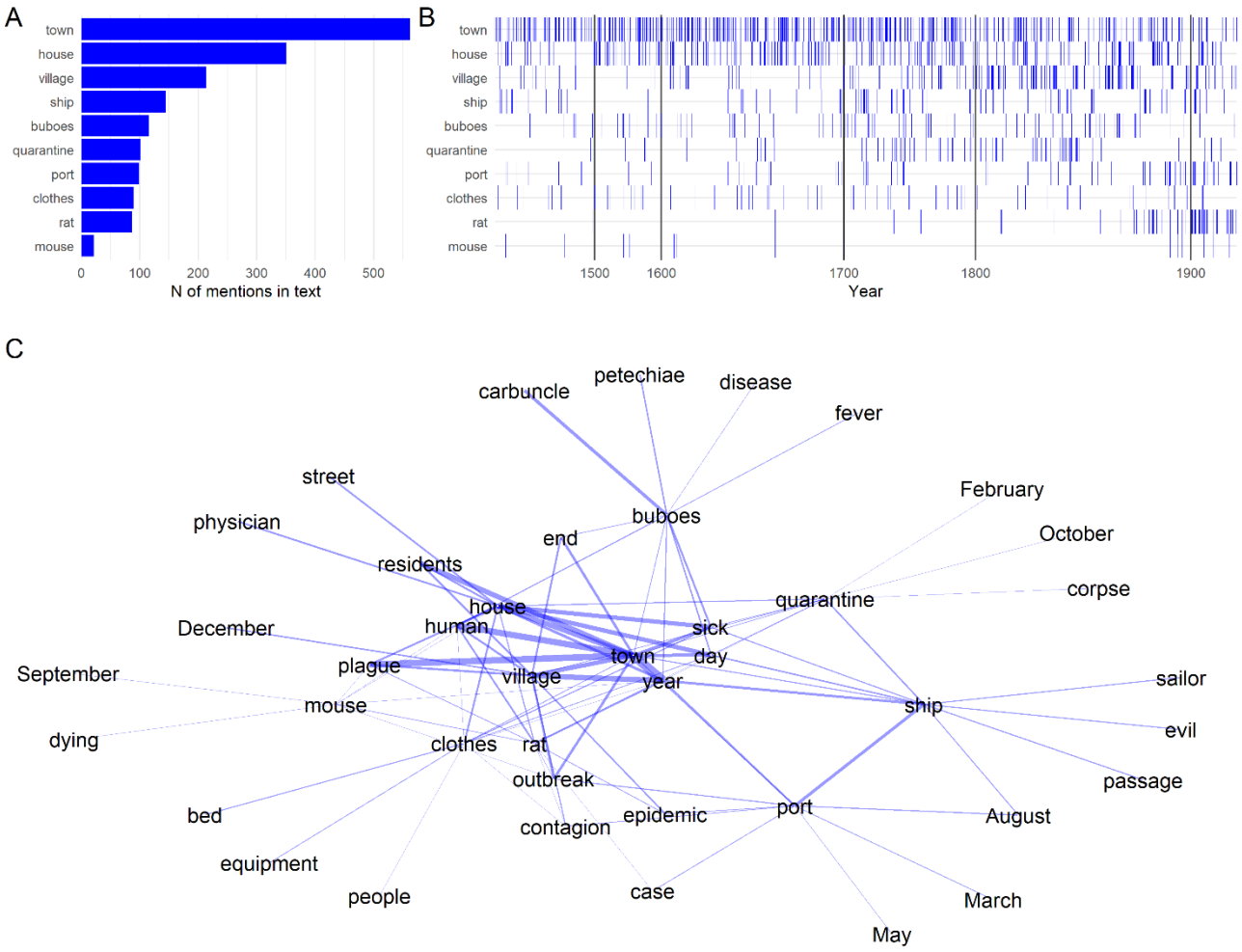

Figure 2. Results from the exploratory text mining. A) Frequency distribution of the most common lemmata related to infrastructure and disease transmission. B) Temporal distribution of the most common lemmata. Each blue line indicates one mention mapped to the corresponding epoch. C) Network plot of the most frequent topics discussed in the book. The width of the blue lines corresponds to the strength of the co-occurrence.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Data description

Comparison of Sticker and Biraben

The final Sticker dataset contained 4474 plague location observations, of which 91.4% could be localized exactly, 8.5% were localized approximatively and 0.1% could not be localized. Of the identified locations, 1631 were unique locations. The Biraben data set had much more data points and unique locations (11,180 observations, 2158 unique locations), of which 95.2% were localized exactly, 3.5% were localized approximately and 1.3% could not be localized) (Table S4). There was some overlap of the data points: 37% of the Sticker data were also in Biraben, and 15% of the Biraben data were also in Sticker. The majority of the data points in Sticker were located in Germany (13.6%), while Biraben had most data points in France (30.2%) (Figure S1A). In both datasets, the majority of locations were places (Sticker 70.1%, Biraben 83.3%). Sticker contained more historical or colloquial regions or administrative units than Biraben, thus the average bounding box diagonal of a location was marginally larger for Sticker (17.6 km vs. 12.1 km) (Figure S3). The most frequent places in Sticker were Istanbul (90 mentions, 2%), London (74 mentions, 1.7%) and Cairo (44 mentions, 1%) (Figure S1B). Biraben listed the most outbreaks for London (166 mentions, 1.5%), Istanbul (118 mentions, 1.1%) and Algiers (114 mentions, 1%). Both data sets have the same overall temporal coverage from the Black Death period to the beginning of the 20th century (Figure S1C). However, the majority of entries in Biraben are from the 16th-17th century, while the majority of Sticker is from the 17th-18th century.

Spatial and temporal extent of second pandemic plague mentions

Figure 3 shows all exactly localized places (without countries, regions or other administrative units) with plague outbreaks reported during the second pandemic (1346-1894) resulting from both datasets. We found 1404 new observations (817 unique locations) in the Sticker dataset, which were not listed in Biraben. These were mainly in eastern Europe, southern Russia and the Caucasus region, as well as India and Iran. London had the largest number of outbreaks (265), followed by Istanbul/Constantinople (205), Algiers (138), Paris (115), Cairo (113), Izmir/Smyrna (107), Venice (103) and Amiens (102). As shown in Figure 4, the spatio-temporal extent of plague mentions shifted considerably over time. Until the 17th century we observe the majority of the data in Central Europe. In the 18th century the focus appears to have shifted to Eastern Europe and North Africa. Finally, in the 19th century the majority of outbreaks seemed to be reported in southeast Europe and West Asia.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

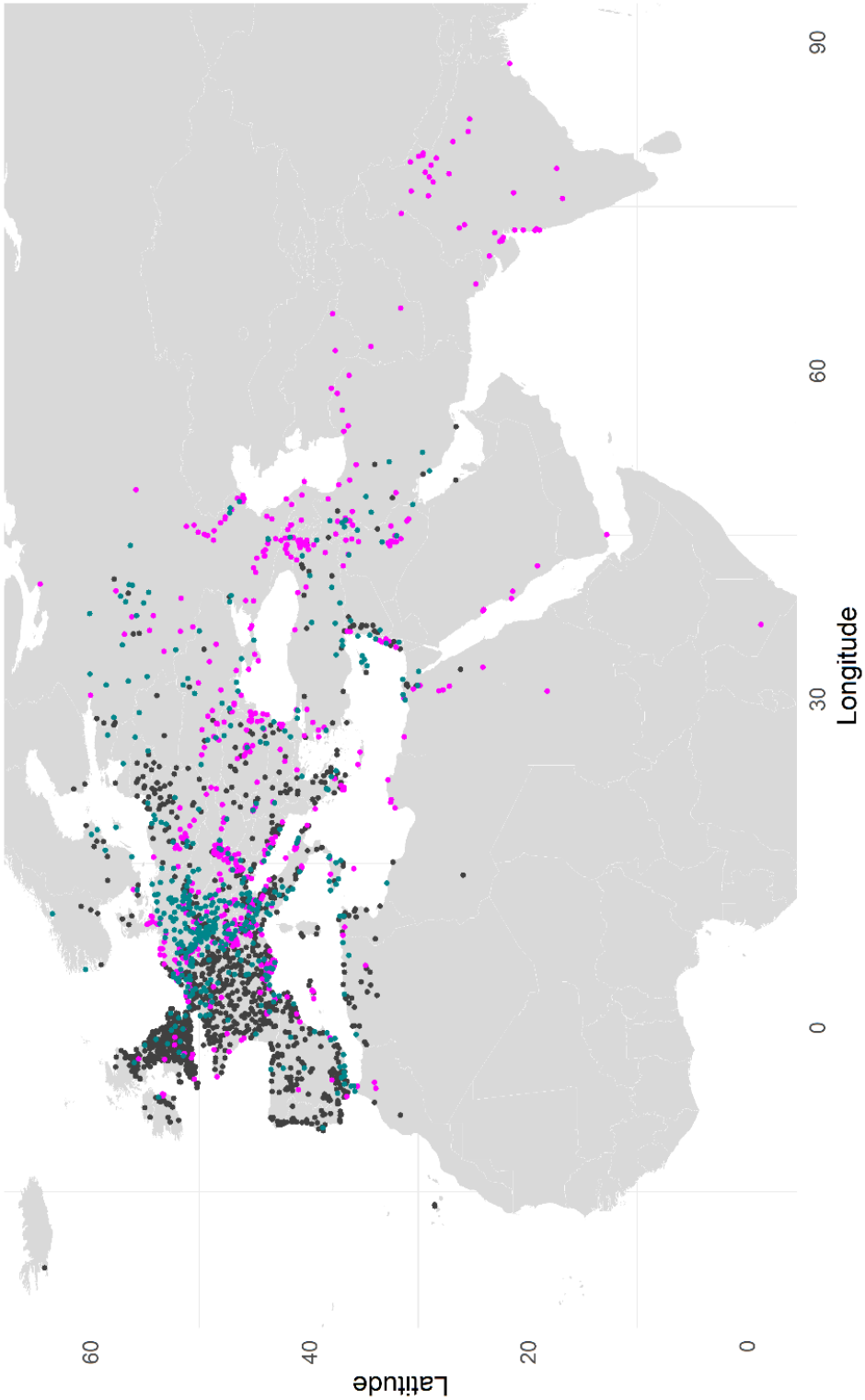

Figure 3. Places with plague during the second pandemic mentioned in the data set of Biraben (grey), Sticker (pink) and both (blue).

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

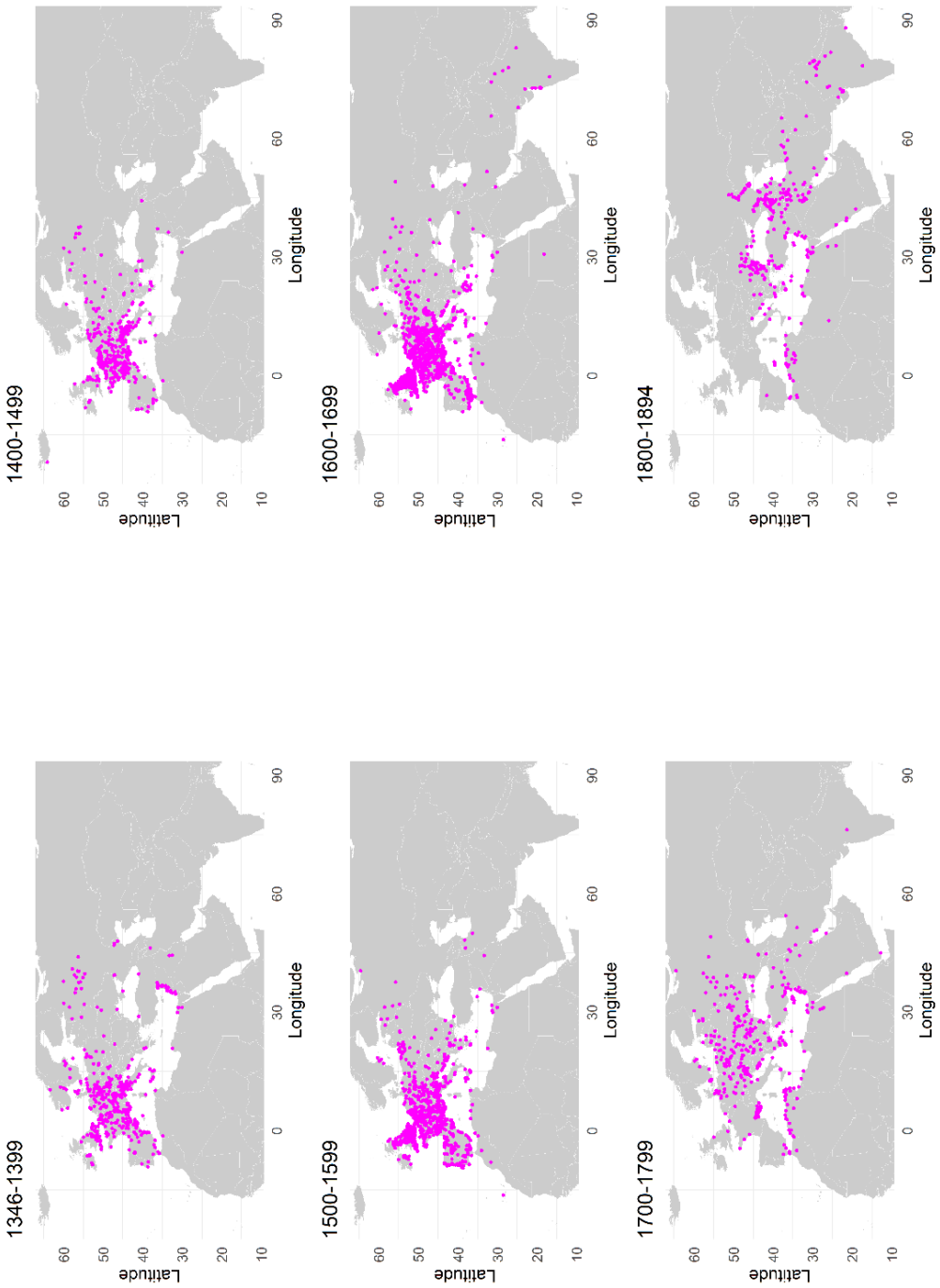

Figure 4. Spatio-temporal extent of the second pandemic plague mentions derived from Biraben and Sticker. The dots denote all exactly localized places, but exclude countries, regions or other administrative areas. For better readability, one data point in Nairobi (Kenya) of an outbreak in 1892 was omitted from the map.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Discussion

Advantages and limitations of NLP and geoparsing

Here we have demonstrated how natural language processing (NLP) libraries and geocoding/geoparsing tools can be used to detect, extract and georeference locations in a running text to facilitate the collection and digitisation of historical plague data. We have shown that the performance of the different algorithms can vary substantially. For the given German text, Stanford's coreNLP and spaCy had a better overall performance than Google's NLP, germaNER and Geoparser.io. While spaCy was better at detecting the true locations (i.e. high sensitivity), Stanford coreNLP was marginally better at avoiding the non-locations (i.e. high specificity). However, all algorithms had a high specificity. Geoparser.io showed a poor performance and missed more than half of the true locations. According to the authors the algorithm works best with English texts, but we couldn't find any information on how the model was trained. Of note, Geoparser.io was discontinued during this study and the website is no longer active. Overall, the sensitivity of all algorithms was imperfect, and a small proportion of locations remained undetected even with the best performing algorithm. All tested algorithms were substantially faster than manual annotation (less than 30 minutes vs. several days per annotator). The sensitivity of Stanford CoreNLP (0.81) and Google NLP (0.78) on Sticker's treatise on plague was comparable to previous results from modern text corpora (0.64-0.89 and 0.77-0.87, respectively) [53-56], but spaCy outperformed its expectations, with a higher sensitivity (0.92) than advertised by the authors of the library (0.85) [43] and estimated in previous studies on English texts (0.57-0.75) [53, 54]. Our F1 score for germaNER was somewhat lower (0.71) than evaluated by the authors of the algorithm (0.81) [44].

NLP libraries combined with geoparsers/geocoding tools are extremely useful to quickly generate quantitative data, but they have some shortcomings when it comes to digitizing plague treatises. As anticipated, these models cannot distinguish whether the mention of a geographical unit is related to a specific plague outbreak or not. This information can only be extracted from the context, but standard models are not trained to recognize these situations. In this study, we have checked the link to a plague outbreak for each location entry manually, which is far from ideal. Moreover, the detection of time units was not optimal. We did not test the year numbers recognition formally, but we observed that Google, spaCy and Stanford CoreNLP don't differentiate between years and any other number. For our gold standard, we used regular expressions (regex), which can identify specific combinations of letters or numbers. The final linking of a specific year with a specific plague location was done manually again, since the order of appearance and the format in which years and locations were reported was not consistent throughout the text. Thus, the current NLP algorithms cannot replace manual work entirely. The decision to use these tools is therefore a trade-off between time gained and precision lost. For larger texts, it may be useful to perform a pilot study on a subset of the text and compare the manual annotation to a NLP approach as we did in our study. If the sensitivity is above an acceptable level and the required additional manual effort is limited, NLP might be a suitable approach. In terms of performance, it is more important to have a high sensitivity than a high specificity, because it is easier to remove false positives in the results than look for false negatives (missed locations) in the text. The main potential (and challenge) of NLP and geoparsing for plague research lies in custom trained models and reproducible, automated workflows. Many of the analysis that we did manually or in separate steps can potentially be improved with an automated procedure. Preprocessing of the raw OCR text prior to applying the NLP algorithms is inevitable, but OCR errors

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

are often consistent and can be corrected with rule-based replacements. Existing NLP or geoparser libraries can be trained specifically on historical texts to improve the recognition of outdated spellings or old place names. The aforementioned plague dot txt project is pioneering the field with its automated OCR optimization and extended NER for the recognition of plague-specific ontology and dates [32]. Given the large body of plague literature, research initiatives like the latter should be encouraged.

Usage and limitations of geo-referenced plague datasets

We here also present two open, georeferenced plague datasets [38]: the newly digitised Sticker dataset and an improved digitisation of Biraben's plague second pandemic appendix. The Biraben dataset has been digitised twice before [57, 58], of which Büntgens version has been used by a number of studies [10-14]. These studies have rightfully drawn criticism for not contextualizing the biases and uncertainties inherent to such aggregated accounts that cover a vast amount of space and time [35, 59]. Both Biraben (and colleagues) as well as Sticker may have been more likely to include sources from specific regions or countries due to easier access to archives or familiarity with the language of the source texts. It is not by accident that the majority of plague mentions of Biraben are in France and the majority of Sticker in Germany. Also, some original sources may not have survived to date due to poor archiving conditions or censorship. Both issues can lead to spatial and/or temporal selection bias in the data. Thus, the absence of plague mentions is not necessarily an absence of outbreaks. Moreover, the retrospective identification of a plague outbreak from historical sources is also often problematic, and the criteria that Sticker and Biraben used to include or exclude information are unclear. In this study, we have not verified the data, but we have provided references to the original sources wherever they were indicated by Sticker, which allows users to cross-check questionable entries. Biraben's treatise was digitized from the tables provided in the appendix, which did not include references for each outbreak. However, the treatise itself includes an extensive bibliography for the origin of the data, which may be linked manually to specific outbreaks. In summary, both datasets have inherent limitations due to the nature of the data collection and the digitisation process. They are presented here as uncommented digitisations of all second and third plague pandemic entries provided in the Biraben and Sticker plague treatises, and should not be regarded as fully prepared and finalized plague data sets. Additional data cleaning is required depending on the research question and type of analysis. For example, the geographical scales of the observations in both datasets are very heterogeneous ranging from small villages to whole countries or historical regions spanning several hundreds of kilometers. For quantitative modelling studies, we recommend to work with data points that represent approximately the same geographic level. We have provided the bounding box coordinates and diagonal for each data point (which gives a rough estimate of the current geographical extent) as well as the type of location, which can be used to select data carefully. We also advise to check for duplicate entries for the same place in the same year, which occurred occasionally when the original dataset listed two separate entries for the same location (for example Saoudje and Boulak for Saoudje-Boulak, presently Mahabad, or individual parishes in London).

In particular quantitative analyses will benefit from improved, georeferenced datasets, for example for the reconstruction of regional transmission chains or potentially the identification of putative historical plague reservoirs [60]. As others have mentioned [35, 61], data collections such as our compilation of Biraben and Sticker can act as a foundation to which more data are added (and faulty

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

data are labelled as such) in order to build an updated database of global plague outbreaks. The growing number of scanned and OCR encoded documents made available online (for example on the Internet Archive) provides a rich resource for historical epidemiology, which should be used with the right tools and the necessary caution. Combining plague data from different sources to fill the spatial and temporal gaps could potentially reduce the problem of spatial and/or temporal representativeness, and improve our understanding of the spatio-temporal spread. Particularly, new data on the plague dissemination in neglected regions such as sub-Saharan Africa [62], Turkey and Southern Asia [62-64] could confirm whether the shift of plague activity from Europe to North Africa in the 16th to 19th century, and the growing presence of plague in Asia in the 17th to 19th century is a real pattern or merely an artefact of missing data in the centuries before. However, consistency in the data definition and collection is crucial. The understanding of the spatio-temporal dynamics of the past and present plague pandemic is a big challenge, which is best tackled with a collaborative and interdisciplinary effort, and in the spirit of open data.

Funding

This work was supported by funding from the Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, and the Research Council of Norway (FRIMEDBIO project 288551).

Ethics statement

Not applicable.

Data accessibility statement

The R code and the digitised plague datasets are available in a public repository [38] (https://doi.org/10.5281/zenodo.4724016).

Competing interests statement

We declare we have no competing interests.

References

[1] Barker, H. 2021 Laying the Corpses to Rest: Grain, Embargoes, and Yersinia pestis in the Black Sea, 1346–48. *Speculum* **96**, 97-126. (doi:10.1086/711596).

[2] Fancy, N. & Green, M.H. 2021 Plague and the Fall of Baghdad (1258). *Medical history* **65**, 157-177. (doi:10.1017/mdh.2021.3).

[3] Bramanti, B., Stenseth, N.C., Walløe, L. & Lei, X. 2016 Plague: A Disease Which Changed the Path of Human Civilization. In: *Yersinia pestis: Retrospective and Perspective* (eds. R. Yang & A. Anisimov), pp. 1-26. Dordrecht, Springer Netherlands.

[4] Hecker, J.F.C. 1865 *Die grossen Volkskrankheiten des Mittelalters: Historisch-pathologische Untersuchungen. Gesammelt und in erweiterter Bearbeitung*. Berlin, Enslin.

[5] Martin, C. 1879 Versuch einer geographischen Darstellung einiger Pestepidemien. *Dr. A. Petermann's Mittheilungen aus Justus Perthes' Geographischer Anstalt* **XXV**.

[6] Creighton, C. 1891 *A history of epidemics in Britain*. 2nd ed. Cambridge, Cambridge University Press.

[7] Dörbeck, F. 1906 *Geschichte der Pestepidemien in Russland von der Gründung des Reiches bis auf die Gegenwart*. Breslau, Kern.

[8] Sticker, G. 1908 *Abhandlungen aus der Seuchengeschichte und Seuchenlehre. Band 1: Die Pest*. Giessen, A. Töpelmann.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

[9] Biraben, J.-N. 1975 *Les hommes et la peste en France et dans les pays européens et méditerranéens*. Paris, Mouton.

[10] Schmid, B.V., Buntgen, U., Easterday, W.R., Ginzler, C., Walloe, L., Bramanti, B. & Stenseth, N.C. 2015 Climate-driven introduction of the Black Death and successive plague reintroductions into Europe. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 3020-3025. (doi:10.1073/pnas.1412887112).

[11] Yue, R.P.H., Lee, H.F. & Wu, C.Y.H. 2017 Trade routes and plague transmission in pre-industrial Europe. *Sci Rep* **7**, 12973. (doi:10.1038/s41598-017-13481-2).

[12] Yue, R.P., Lee, H.F. & Wu, C.Y. 2016 Navigable rivers facilitated the spread and recurrence of plague in pre-industrial Europe. *Sci Rep* **6**, 34867. (doi:10.1038/srep34867).

[13] Yue, R.P.H. & Lee, H.F. 2018 Pre-industrial plague transmission is mediated by the synergistic effect of temperature and aridity index. *BMC infectious diseases* **18**, 134. (doi:10.1186/s12879-018-3045-5).

[14] Yue, R.P.H. & Lee, H.F. 2020 Drought-induced spatio-temporal synchrony of plague outbreak in Europe. *Sci Total Environ* **698**, 134138. (doi:10.1016/j.scitotenv.2019.134138).

[15] Marien, G. 2009 The Black Death in early Ottoman territories: 1347–1550 [PhD dissertation], Bilkent University, Ankara, Turkey.

[16] Varlik, N. 2015 *Plague and Empire in the Early Modern Mediterranean World*. Cambridge, Cambridge University Press.

[17] Panzac, D. 1985 *La peste dans l'empire ottoman, 1700-1850*, Peeters.

[18] Noordegraaf, L. & Valk, G. 1996 *De gave Gods: de pest in Holland vanaf de late middeleeuwen*, B. Bakker.

[19] Roosen, J. & Curtis, D.R. 2019 The 'light touch' of the Black Death in the Southern Netherlands: an urban trick? *Econ Hist Rev* **72**, 32-56. (doi:10.1111/ehr.12667).

[20] Dols, M.W. 1977 *The "black death" in the Middle East*. Princeton (N.J.), Princeton University Press.

[21] Dols, M.W. 1979 The second plague pandemic and its recurrences in the Middle East: 1347-1894. *J Econ Soc Hist Orient* **22**, 162-189.

[22] Slavin, P. 2021 Out of the West: Formation of a Permanent Plague Reservoir in South-Central Germany (1349–1356) and its Implications∗. *Past & Present*. (doi:10.1093/pastj/gtaa028).

[23] Frandsen, K.E. 2010 *The Last Plague in the Baltic Region, 1709-1713*, Museum Tusculanum Press; 537 p.

[24] Panzac, D. 2004 La peste dans les possessions insulaires du Grand Seigneur (xviie-xixe siècles). In: *Insularités ottomanes* (eds. N. Vatin & G. Veinstein), pp. 223-240, Institut français d'études anatoliennes.

[25] Fazlinejad, A. & Ahmadi, F. 2018 The Black Death in Iran, according to Iranian Historical Accounts from the Fourteenth through Fifteenth Centuries. *Journal of Persianate Studies* **11**, 56-71. (doi:10.1163/18747167-12341321).

[26] Salathe, M. 2018 Digital epidemiology: what is it, and where is it going? *Life Sci Soc Policy* **14**, 1. (doi:10.1186/s40504-017-0065-7).

[27] Dreisbach, C., Koleck, T.A., Bourne, P.E. & Bakken, S. 2019 A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform* **125**, 37-46. (doi:10.1016/j.ijmedinf.2019.02.008).

[28] Broniatowski, D.A., Paul, M.J. & Dredze, M. 2013 National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one* **8**, e83672. (doi:10.1371/journal.pone.0083672).

[29] Won, M., Murrieta-Flores, P. & Martins, B. 2018 Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities* **5**. (doi:10.3389/fdigh.2018.00002).

[30] Clifford, J., Alex, B., Coates, C.M., Klein, E. & Watson, A. 2016 Geoparsing history: Locating commodities in ten million pages of nineteenth-century sources. *Historical Methods: A*

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Journal of Quantitative and Interdisciplinary History **49**, 115-131. (doi:10.1080/01615440.2015.1116419).

[31] Murrieta-Flores, P., Baron, A., Gregory, I., Hardie, A. & Rayson, P. 2015 Automatically Analyzing Large Texts in a GIS Environment: The Registrar General's Reports and Cholera in the 19th Century. *Transactions in GIS* **19**, 296-320. (doi:10.1111/tgis.12106).

[32] Casey, A., Bennett, M., Tobin, R., Grover, C., Walker, I., Engelmann, L. & Alex, B. 2020 Plague Dot Text: Text mining and annotation of outbreak reports of the Third Plague Pandemic (1894-1952). *Journal of Data Mining & Digital Humanities*.

[33] Gritta, M., Pilehvar, M.T., Limsopatham, N. & Collier, N. 2018 What's missing in geographical parsing? *Lang Resour Eval* **52**, 603-623. (doi:10.1007/s10579-017-9385-8).

[34] Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S. & Ball, J. 2010 Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philos Trans A Math Phys Eng Sci* **368**, 3875-3889. (doi:10.1098/rsta.2010.0149).

[35] van Bavel, B.J.P., Curtis, D.R., Hannaford, M.J., Moatsos, M., Roosen, J. & Soens, T. 2019 Climate and society in long-term perspective: Opportunities and pitfalls in the use of historical datasets. *Wiley Interdiscip Rev Clim Change* **10**, e611. (doi:10.1002/wcc.611).

[36] Green, M. & Roosen, J. Biraben 2.0: A Black Death Digital Archive. Available from: https://www.radcliffe.harvard.edu/seminars-and-workshops/biraben-black-death-digital-archive. [Accessed Oct 16, 2020].

[37] Eckart de Castilho, R., Újdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A. & Biemann, C.2016.A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures.

[38] Krauer, F. & Schmid, B.V. 2021 Datasets and code for "Mapping the plague through natural language processing". (doi:10.5281/zenodo.4724016).

[39] Google Ireland Limited. Google Cloud Natural Language API. Available from: https://cloud.google.com/natural-language/docs/?hl=en. [Accessed Oct 16, 2020].

[40] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. & McClosky, D.2014.The Stanford CoreNLP Natural Language Processing Toolkit.

[41] Faruqui, M. & Padó, S.2010.Training and evaluating a German named entity recognizer with semantic generalization.

[42] Explosion. spaCy v2.x. Available from: https://spacy.io/. [Accessed Jul 13, 2019].

[43] Explosion. 2019 de_core_news_sm-2.1.0. (

[44] Benikova, D., Yimam, S.M., Santhanam, P. & Biemann, C.2015.GermaNER: Free Open German Named Entity Recognition Tool.

[45] Geoparser Inc. geoparser.io. Available from: https://geoparser.io/. [Accessed Jul 8, 2019].

[46] Arnold, T. & Tilton, L. coreNLP: Wrappers Around Stanford CoreNLP Tools. Available from: https://CRAN.R-project.org/package=coreNLP. [Accessed Oct 10, 2019].

[47] Benoit, K. & Matsuo, A. spacyr: Wrapper to the 'spaCy' 'NLP' Library. Available from: https://CRAN.R-project.org/package=spacyr. [Accessed Oct 10, 2019].

[48] GeoNames. Available from: http://www.geonames.org. [Accessed Nov 11, 2019].

[49] Naber, D. Deutsches Morphologie-Lexikon. Available from: http://danielnaber.de/morphologie/. [Accessed Apr 15, 2019].

[50] Simond PL. 1898 La propagation de la peste. *Annales de l'Institut Pasteur* **12**, 625–687.

[51] Wood, W. 1859 *The History and Antiquities of Eyam: With a Minute Account of the Great Plague which Desolated that Village in the Year 1666*, Bell and Daldy.

[52] Dean, K.R., Krauer, F., Walloe, L., Lingjaerde, O.C., Bramanti, B., Stenseth, N.C. & Schmid, B.V. 2018 Human ectoparasites and the spread of plague in Europe during the Second Pandemic. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 1304-1309. (doi:10.1073/pnas.1715640115).

[53] Gritta, M., Pilehvar, M.T. & Collier, N. 2020 A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics. *Lang Resour Eval* **54**, 683-712. (doi:10.1007/s10579-019-09475-3).

medRxiv preprint doi: https://doi.org/10.1101/2021.04.27.21256212; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

[54] Schmitt, X., Kubler, S., Robert, J., Papadakis, M. & LeTraon, Y.2019.A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate.

[55] Pinto, A., Gonçalo Oliveira, H. & Oliveira Alves, A.2016.Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text.

[56] Dale, R. 2018 Text analytics APIs, Part 1: The bigger players. *Natural Language Engineering* **24**, 317-324. (doi:10.1017/s1351324918000013).

[57] Buntgen, U., Ginzler, C., Esper, J., Tegel, W. & McMichael, A.J. 2012 Digitizing historical plague. *Clin Infect Dis* **55**, 1586-1588. (doi:10.1093/cid/cis723).

[58] Atanasiu V, Priol C, Tournieroux A & E, O. Georeferences for places of plague occurrence in Europe 1347-1600. Available from: https://bernstein.oeaw.ac.at/atlas/yersinia-description.pdf. [Accessed Mar 3, 2019].

[59] Roosen, J. & Curtis, D.R. 2018 Dangers of Noncritical Use of Historical Plague Data. *Emerging infectious diseases* **24**, 103-110. (doi:10.3201/eid2401.170477).

[60] Carmichael, A.G. 2014 Plague persistance in Western Europe: a hypothesis. *The Medieval Globe* **1**, 157-191.

[61] Benedictow, O.J. 2019 Biraben's lists of the plague epidemics of the second plague pandemic, 1346 - c. 1690: problems, basis, uses. *Annales de démographie historique* **n°138**, 213-223. (doi:10.3917/adh.138.0213).

[62] Green, M.H. 2018 Putting Africa on the Black Death map: Narratives from genetics and history. *Afriques*. (doi:10.4000/afriques.2125).

[63] Varlik, N. 2020 The plague that never left: restoring the Second Pandemic to Ottoman and Turkish history in the time of COVID-19. *New Perspectives on Turkey*, 1-14. (doi:10.1017/npt.2020.27).

[64] Green, M. 2014 Taking "Pandemic" Seriously: Making the Black Death Global. *Medieval Globe* **1**, 27-61.