

1 Mapping the plague through natural language processing

2 Fabienne Krauer^{1,2}, Boris V. Schmid¹

3
4 ¹*Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo,*
5 *NO-0316 Oslo, Norway*

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

9 Abstract

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

27 Introduction

28 The dissemination of historical plague across the globe is being studied and discussed since many
29 decades. The second pandemic entered Europe in 1347 [1], but it may have started a century earlier
30 in Central Asia and China [2]. The pandemic caused millions of casualties before gradually
31 disappearing from Europe and the Mediterranean in the 18th century. The third pandemic is thought
32 to have started towards the end of the 18th century in Yunnan (China) [3]. It reached Hong Kong in
33 1894 from where it spread globally. Since the 19th century - and perhaps even earlier - many scholars
34 have collected data on plague outbreaks in a more or less systematic manner. Among the earliest
35 compilations with a broader geographical coverage are the works of Hecker in 1865 [4], Martin in
36 1879 [5], Creighton in 1891 for the British Isles [6], Dörbeck in 1906 for Russia [7] or Sticker in
37 1908 for the whole world [8]. In 1975, Biraben combined data from some of these and many other
38 sources into the largest tabellaric collection of yearly plague outbreaks published to date [9]. His list
39 has been used by many for spatio-temporal analyses of plague dispersion in Europe [10-14]. In the
40 decades after Biraben, many more scholars have continued his legacy and collected, reviewed and
41 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.

42 Panzac [17] for the Ottoman Empire, Noordegraaf [18] and Roosen [19] for the Low Countries, Dols
43 for the Black Death in the Middle East [20, 21], Slavín for the *pestis secundis* [22], Frandsen for the
44 1709-13 plague in the Baltics [23], Panzac for the islands in the eastern Mediterranean in the 17th to
45 19th century [24] or Fazlinejad and Ahmadi for 14th and 15th century Iran [25]. Most of these
46 publications (both recent or more historical) contain information about the places and times of
47 epidemics, which are the core aspects of infectious disease epidemiology. Many offer also additional
48 information such as the size of the outbreaks, symptoms, treatments, control measures, putative
49 transmission routes and ecological aspects. The body of publications is large, but narrative text must
50 be converted into quantitative data in order to be usable for epidemiological analyses. However, the
51 extraction of data from running text is time and labour intensive.

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

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

77 In general, text mining tools can accelerate the generation of large datasets, but their performance has
78 to be sufficient to outweigh the errors arising from the automated process. The performance of these
79 algorithms depends on the chosen model or algorithm, and the structure and language of the text.
80 Ideally, an NLP algorithm has a high recall or sensitivity (e. g. the proportion of locations that are
81 correctly identified as locations) and a high specificity (e. g. the proportion of non-locations that are
82 correctly identified as non-locations). Various NLP algorithms and libraries have been tested for
83 modern English medical and non-medical texts and their performances differ substantially (see e.g.
84 [27, 33]). The literature on performance evaluation of NLP libraries for more historical texts is sparser.

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

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

99 **Methods**

100 **Source text**

101 Our source text is a German plague treatise (“Geschichte der Pest”) published in 1908 by Georg
102 Sticker [8]. Sticker was a German physician, who together with Robert Koch was sent to Bombay to
103 investigate the plague epidemic in 1897. In the first part of the book, he narrates the historical spread
104 of plague on five continents (Europe, Asia, the Americas, Africa and Australia) in chronological order
105 from biblical times to 1908. In the second part, Sticker elaborates on plague ecology, microbiology
106 and pathology. For the historical part, he combined information from secondary literature, but he also
107 consulted original sources (see his introduction). It is divided in 16 Chapters, of which the first four
108 chapters are about the first pandemic, chapters 5 to 15 are about the second pandemic and the chapters
109 15 and 16 refer to the third pandemic. For our analysis, we focus on plague during the second and
110 third pandemic, corresponding to the chapters five to sixteen of Sticker’s treatise on plague (pages 42
111 to 399). The structure of the text is a combination of running text interspersed with semi-tabulated
112 year and place listings. The running text contains both specific information about places that mention
113 plague in a given year, but also general information on plague as well as historical anecdotes and
114 elaborations. A scanned OCR version of the book is freely available on the Internet Archive
115 (<https://archive.org/details/abhandlungenausd01stic/mode/2up>).

116 **Preprocessing and establishment of gold standard**

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

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

158 **Toponym tagging performance evaluation**

159 We tested four different NLP libraries and one geoparser for the tagging of toponyms: Google NLP
160 [39], Stanford CoreNLP [40] with the pre-trained German model version 2018-10-05 [41], spaCy [42]
161 with the pre-trained German model version 2.1.0 [43], germaNER [44] and Geoparser.io [45]. For a
162 technical comparison of the libraries see supplement Table S2. We performed syntax analysis (POS)
163 and entity recognition (NER). Google NLP differentiates between person, location, organization,
164 event, work of art, consumer good and other, while Stanford CoreNLP differentiates between person,
165 location, organization, misc and “o” (outside, not classified). The german model used with SpaCy as
166 well as germaNER both differentiate between location, person, organization and misc/other.
167 Geoparser.io only returns toponyms and the corresponding GIS information but not the tokenization
168 of the complete text. All algorithms accept running text except germaNER, which requires a priori
169 tokenization. We therefore used the tokenization returned by spaCy as an input for germaNER. The

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

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

201 **Geocoding performance evaluation**

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

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

224 **Exploratory text mining**

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

240 **Data description**

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

248

249 **Results**

250 **Gold standard**

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

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

262

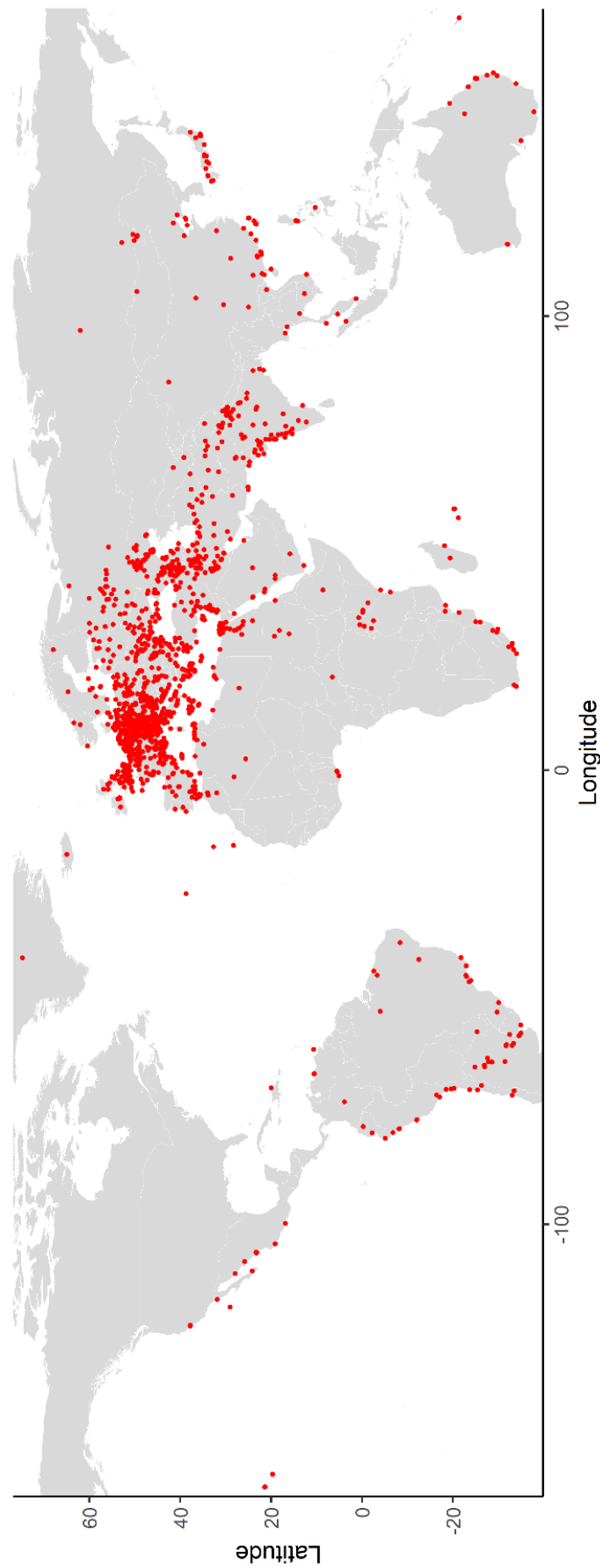


Figure 1. Spatial coverage of all geocoded geographical units (exact and approximate). Note that the data include different administrative levels from village to country, and the dots denote the centroids of each geographical entity.

263 **Toponym tagging performance evaluation**

264 The spaCy and the Stanford CoreNLP tokenizers yielded a similar number of total tokens (146,766
265 and 146,743 respectively) while Google NLP returned less tokens (146,340) (Table S3). GermaNER
266 identified the most entities (34% of all recognized tokens = 50,374), followed by Google NLP (23%
267 of all recognized tokens = 33,925), spaCy (9% of all recognized tokens = 12,963), Stanford coreNLP
268 (6.5% of all recognized tokens = 9522) and Geoparser.io (3563). We then used the tokenization
269 returned by the spaCy algorithm as the base and mapped all results of the NER analysis onto the
270 tokens to compare the different approaches. With the gold standard, 5.7% of all tokens were classified
271 as locations. The text mining algorithms of Google, Stanford, spaCy and germaNER resulted in 6.5%,
272 4.8%, 6.9% and 4.0% of the tokens categorized as locations. The Geoparser.io algorithm mapped
273 only 2.5% of the tokens as locations.

274 We then evaluated the performance of the five text mining algorithms compared to the Gold standard
275 for the detection of locations (Table 1). Overall, the proportion of correctly identified entities
276 (accuracy) was large for all five libraries (range 0.96-0.99). The spaCy library showed the best
277 sensitivity (0.92), followed by Stanford CoreNLP (0.81), Google (0.78), germanNER (0.61) and
278 Geoparser.io (0.41). The specificity was equally high for all algorithms (range 0.98-0.99). Stanford
279 coreNLP had the highest precision (PPV, 0.95, i.e. 5% false positives.), followed by Geoparser.io
280 (0.91), germaNER (0.86), spaCy (0.76) and Google (0.67). The F1 scores and Cohen's kappa
281 coefficients suggested a good overall performance for Stanford CoreNLP (0.87 and 0.87) and spaCy
282 (0.83 and 0.82), a mediocre overall performance for Google NLP (0.72 and 0.70) and GermanER
283 (0.71 and 0.70) and a poor performance for Geoparser.io (0.56 and 0.55). Of note, the slightly higher
284 false positive rate of Google NLP can be explained by its rather broad definition of "location", which
285 includes not only geographical entities but also nouns related to physical locations such as "Stadt"
286 (town, city), "Provinz" (province) or "Haus" (house). Of the 8313 location tokens, only 26% were
287 correctly identified by all algorithms. Locations that were missed by all algorithms included
288 Germanized spelling (e.g. "Hoschiarpur" for "Hoshiarpur"), latin spelling (e.g. "Centumcellae" for
289 "Civitavecchia"), composite entities (e.g. "Gurjewscher Kreis"), historic regions (e.g. "Podolien") or
290 ambiguous words (e.g. "Sind" is a location but also a conjugated verb form of "to be"). The only
291 token that was falsely identified as a location by all algorithms was "Santa Maria", which can be both
292 a church name as well as a place name.

293

294

295

296

297

298

299

300

301 **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

302

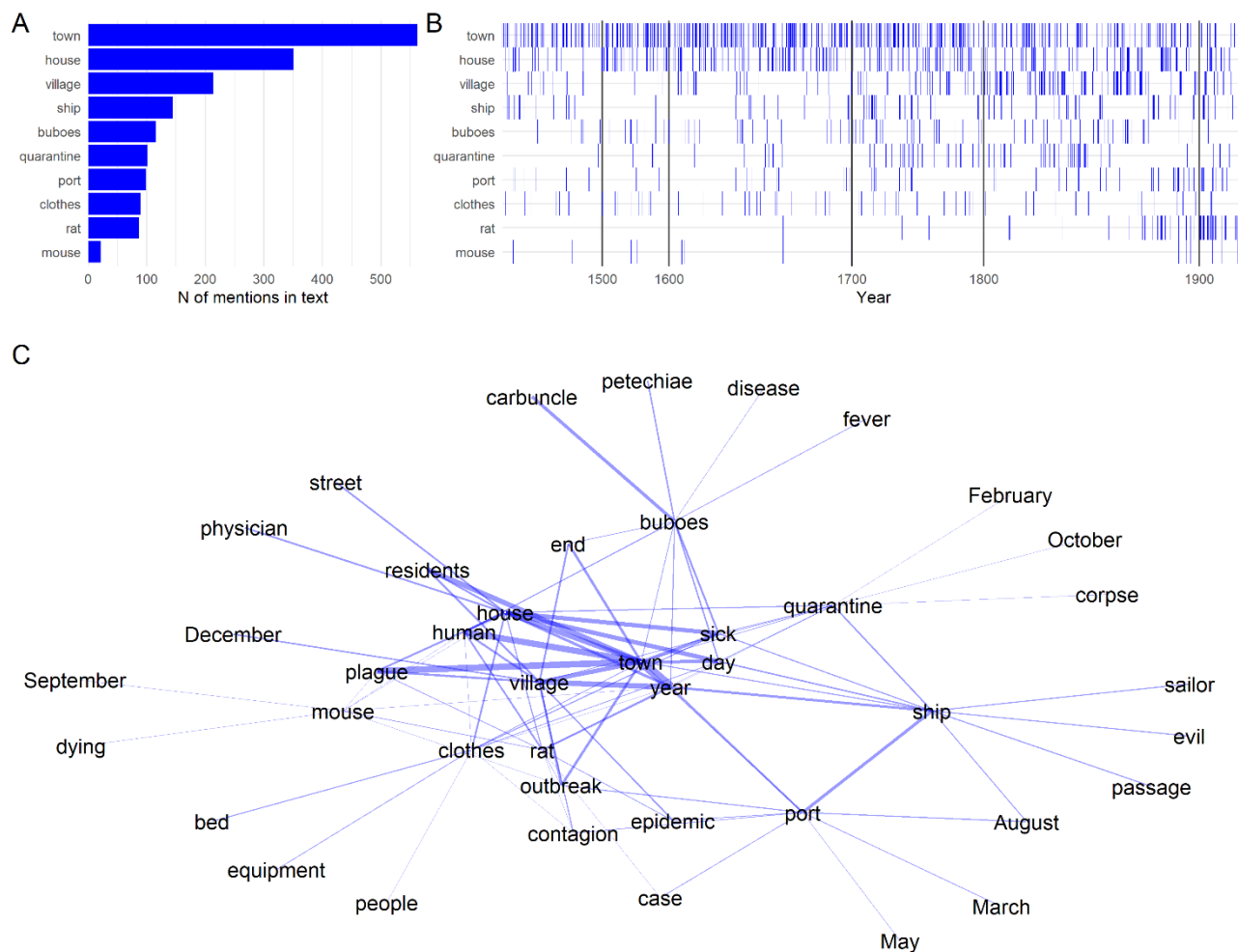
303 **Geocoding performance evaluation**

304 To evaluate the geocoding performances, we compared the 4087 locations from the gold standard
305 that were true plague locations and that were identifiable with the exact coordinates. Geonames
306 identified substantially more toponyms (86.6 %) than Geoparser.io (50.1%). However, both
307 geocoding services performed similarly well in the correct identification of a location (Geoparser.io
308 85.2% and Geonames 82.2%). Of note, the performance could be improved substantially if additional
309 information was provided such as the country or region. However, for this study we evaluated the
310 situation where there is no additional information available and where the performance depends on
311 the best guess of the Geocoding algorithm. Many of the mismatches occurred for regions where places
312 were renamed as the ruling power changed, through colonization or the contraction and expansion of
313 empires.

314 **Exploratory text mining**

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

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



335

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

341

342 **Data description**

343 **Comparison of Sticker and Biraben**

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

361 **Spatial and temporal extent of second pandemic plague mentions**

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

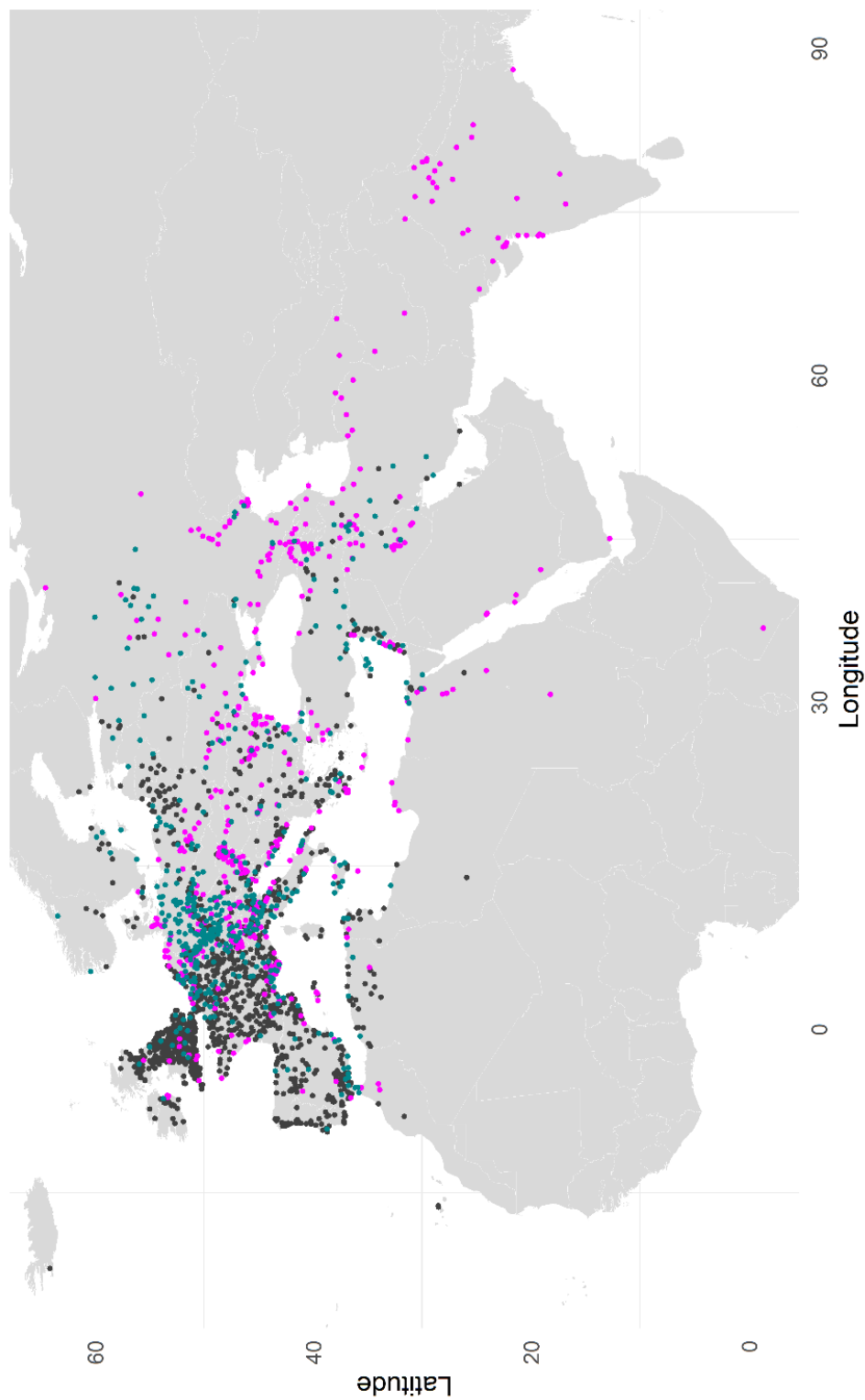


Figure 3. Places with plague during the second pandemic mentioned in the data set of Biraben (grey), Sticker (pink) and both (blue).

372

373

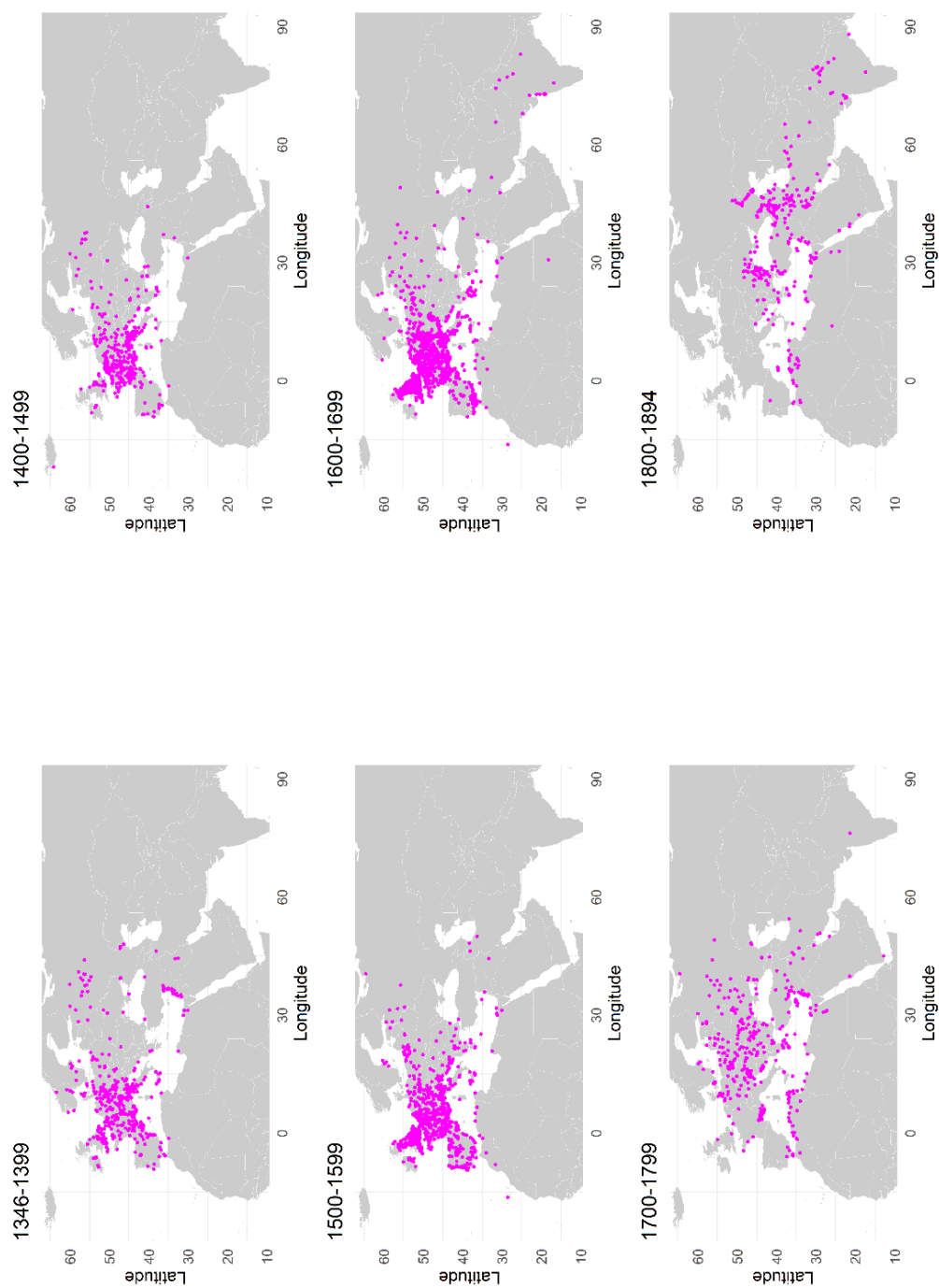


Figure 4. Spatio-temporal extent of the second pandemic plague mentions derived from Biraben and Sticker. The dots denote all exactly localized places, but exclude countries, regions or other administrative areas. For better readability, one data point in Nairobi (Kenya) of an outbreak in 1892 was omitted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

374

375

376 Discussion

377 Advantages and limitations of NLP and geoparsing

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

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

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

425 **Usage and limitations of geo-referenced plague datasets**

426 We here also present two open, georeferenced plague datasets [38]: the newly digitised Sticker dataset
427 and an improved digitisation of Biraben's plague second pandemic appendix. The Biraben dataset
428 has been digitised twice before [57, 58], of which Büntgens version has been used by a number of
429 studies [10-14]. These studies have rightfully drawn criticism for not contextualizing the biases and
430 uncertainties inherent to such aggregated accounts that cover a vast amount of space and time [35,
431 59]. Both Biraben (and colleagues) as well as Sticker may have been more likely to include sources
432 from specific regions or countries due to easier access to archives or familiarity with the language of
433 the source texts. It is not by accident that the majority of plague mentions of Biraben are in France
434 and the majority of Sticker in Germany. Also, some original sources may not have survived to date
435 due to poor archiving conditions or censorship. Both issues can lead to spatial and/or temporal
436 selection bias in the data. Thus, the absence of plague mentions is not necessarily an absence of
437 outbreaks. Moreover, the retrospective identification of a plague outbreak from historical sources is
438 also often problematic, and the criteria that Sticker and Biraben used to include or exclude information
439 are unclear. In this study, we have not verified the data, but we have provided references to the
440 original sources wherever they were indicated by Sticker, which allows users to cross-check
441 questionable entries. Biraben's treatise was digitized from the tables provided in the appendix, which
442 did not include references for each outbreak. However, the treatise itself includes an extensive
443 bibliography for the origin of the data, which may be linked manually to specific outbreaks. In
444 summary, both datasets have inherent limitations due to the nature of the data collection and the
445 digitisation process. They are presented here as uncommented digitisations of all second and third
446 plague pandemic entries provided in the Biraben and Sticker plague treatises, and should not be
447 regarded as fully prepared and finalized plague data sets. Additional data cleaning is required
448 depending on the research question and type of analysis. For example, the geographical scales of the
449 observations in both datasets are very heterogeneous ranging from small villages to whole countries
450 or historical regions spanning several hundreds of kilometers. For quantitative modelling studies, we
451 recommend to work with data points that represent approximately the same geographic level. We
452 have provided the bounding box coordinates and diagonal for each data point (which gives a rough
453 estimate of the current geographical extent) as well as the type of location, which can be used to select
454 data carefully. We also advise to check for duplicate entries for the same place in the same year,
455 which occurred occasionally when the original dataset listed two separate entries for the same location
456 (for example Saoudje and Boulak for Saoudje-Boulak, presently Mahabad, or individual parishes in
457 London).

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

462 data are labelled as such) in order to build an updated database of global plague outbreaks. The
463 growing number of scanned and OCR encoded documents made available online (for example on the
464 Internet Archive) provides a rich resource for historical epidemiology, which should be used with the
465 right tools and the necessary caution. Combining plague data from different sources to fill the spatial
466 and temporal gaps could potentially reduce the problem of spatial and/or temporal representativeness,
467 and improve our understanding of the spatio-temporal spread. Particularly, new data on the plague
468 dissemination in neglected regions such as sub-Saharan Africa [62], Turkey and Southern Asia [62-
469 64] could confirm whether the shift of plague activity from Europe to North Africa in the 16th to 19th
470 century, and the growing presence of plague in Asia in the 17th to 19th century is a real pattern or
471 merely an artefact of missing data in the centuries before. However, consistency in the data definition
472 and collection is crucial. The understanding of the spatio-temporal dynamics of the past and present
473 plague pandemic is a big challenge, which is best tackled with a collaborative and interdisciplinary
474 effort, and in the spirit of open data.

475 **Funding**

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

478 **Ethics statement**

479 Not applicable.

480 **Data accessibility statement**

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

483 **Competing interests statement**

484 We declare we have no competing interests.

485 **References**

- 486 [1] Barker, H. 2021 Laying the Corpses to Rest: Grain, Embargoes, and *Yersinia pestis* in the Black
487 Sea, 1346–48. *Speculum* **96**, 97-126. (doi:10.1086/711596).
488 [2] Fancy, N. & Green, M.H. 2021 Plague and the Fall of Baghdad (1258). *Medical history* **65**, 157-
489 177. (doi:10.1017/mdh.2021.3).
490 [3] Bramanti, B., Stenseth, N.C., Walløe, L. & Lei, X. 2016 Plague: A Disease Which Changed the
491 Path of Human Civilization. In: *Yersinia pestis: Retrospective and Perspective* (eds. R. Yang
492 & A. Anisimov), pp. 1-26. Dordrecht, Springer Netherlands.
493 [4] Hecker, J.F.C. 1865 *Die grossen Volkskrankheiten des Mittelalters: Historisch-pathologische*
494 *Untersuchungen. Gesammelt und in erweiterter Bearbeitung*. Berlin, Enslin.
495 [5] Martin, C. 1879 Versuch einer geographischen Darstellung einiger Pestepidemien. *Dr. A.*
496 *Petermann's Mittheilungen aus Justus Perthes' Geographischer Anstalt* **XXV**.
497 [6] Creighton, C. 1891 *A history of epidemics in Britain*. 2nd ed. Cambridge, Cambridge University
498 Press.
499 [7] Dörbeck, F. 1906 *Geschichte der Pestepidemien in Russland von der Gründung des Reiches bis*
500 *auf die Gegenwart*. Breslau, Kern.
501 [8] Sticker, G. 1908 *Abhandlungen aus der Seuchengeschichte und Seuchenlehre. Band 1: Die Pest*.
502 Giessen, A. Töpelmann.

- 503 [9] Biraben, J.-N. 1975 *Les hommes et la peste en France et dans les pays européens et*
504 *méditerranéens*. Paris, Mouton.
- 505 [10] Schmid, B.V., Buntgen, U., Easterday, W.R., Ginzler, C., Walloe, L., Bramanti, B. & Stenseth,
506 N.C. 2015 Climate-driven introduction of the Black Death and successive plague
507 reintroductions into Europe. *Proceedings of the National Academy of Sciences of the United*
508 *States of America* **112**, 3020-3025. (doi:10.1073/pnas.1412887112).
- 509 [11] Yue, R.P.H., Lee, H.F. & Wu, C.Y.H. 2017 Trade routes and plague transmission in pre-
510 industrial Europe. *Sci Rep* **7**, 12973. (doi:10.1038/s41598-017-13481-2).
- 511 [12] Yue, R.P., Lee, H.F. & Wu, C.Y. 2016 Navigable rivers facilitated the spread and recurrence of
512 plague in pre-industrial Europe. *Sci Rep* **6**, 34867. (doi:10.1038/srep34867).
- 513 [13] Yue, R.P.H. & Lee, H.F. 2018 Pre-industrial plague transmission is mediated by the synergistic
514 effect of temperature and aridity index. *BMC infectious diseases* **18**, 134.
515 (doi:10.1186/s12879-018-3045-5).
- 516 [14] Yue, R.P.H. & Lee, H.F. 2020 Drought-induced spatio-temporal synchrony of plague outbreak
517 in Europe. *Sci Total Environ* **698**, 134138. (doi:10.1016/j.scitotenv.2019.134138).
- 518 [15] Marien, G. 2009 *The Black Death in early Ottoman territories: 1347–1550* [PhD dissertation],
519 Bilkent University, Ankara, Turkey.
- 520 [16] Varlik, N. 2015 *Plague and Empire in the Early Modern Mediterranean World*. Cambridge,
521 Cambridge University Press.
- 522 [17] Panzac, D. 1985 *La peste dans l'empire ottoman, 1700-1850*, Peeters.
- 523 [18] Noordegraaf, L. & Valk, G. 1996 *De gave Gods: de pest in Holland vanaf de late middeleeuwen*,
524 B. Bakker.
- 525 [19] Roosen, J. & Curtis, D.R. 2019 The 'light touch' of the Black Death in the Southern Netherlands:
526 an urban trick? *Econ Hist Rev* **72**, 32-56. (doi:10.1111/ehr.12667).
- 527 [20] Dols, M.W. 1977 *The "black death" in the Middle East*. Princeton (N.J.), Princeton University
528 Press.
- 529 [21] Dols, M.W. 1979 The second plague pandemic and its recurrences in the Middle East: 1347-
530 1894. *J Econ Soc Hist Orient* **22**, 162-189.
- 531 [22] Slavin, P. 2021 Out of the West: Formation of a Permanent Plague Reservoir in South-Central
532 Germany (1349–1356) and its Implications*. *Past & Present*. (doi:10.1093/pastj/gtaa028).
- 533 [23] Frandsen, K.E. 2010 *The Last Plague in the Baltic Region, 1709-1713*, Museum Tusulanum
534 Press; 537 p.
- 535 [24] Panzac, D. 2004 La peste dans les possessions insulaires du Grand Seigneur (xviiie-xixe siècles).
536 In: *Insularités ottomanes* (eds. N. Vatin & G. Veinstein), pp. 223-240, Institut français
537 d'études anatoliennes.
- 538 [25] Fazlinejad, A. & Ahmadi, F. 2018 The Black Death in Iran, according to Iranian Historical
539 Accounts from the Fourteenth through Fifteenth Centuries. *Journal of Persianate Studies* **11**,
540 56-71. (doi:10.1163/18747167-12341321).
- 541 [26] Salathe, M. 2018 Digital epidemiology: what is it, and where is it going? *Life Sci Soc Policy* **14**,
542 1. (doi:10.1186/s40504-017-0065-7).
- 543 [27] Dreisbach, C., Koleck, T.A., Bourne, P.E. & Bakken, S. 2019 A systematic review of natural
544 language processing and text mining of symptoms from electronic patient-authored text data.
545 *Int J Med Inform* **125**, 37-46. (doi:10.1016/j.ijmedinf.2019.02.008).
- 546 [28] Broniatowski, D.A., Paul, M.J. & Dredze, M. 2013 National and local influenza surveillance
547 through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one* **8**, e83672.
548 (doi:10.1371/journal.pone.0083672).
- 549 [29] Won, M., Murrieta-Flores, P. & Martins, B. 2018 Ensemble Named Entity Recognition (NER):
550 Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers*
551 *in Digital Humanities* **5**. (doi:10.3389/fdigh.2018.00002).
- 552 [30] Clifford, J., Alex, B., Coates, C.M., Klein, E. & Watson, A. 2016 Geoparsing history: Locating
553 commodities in ten million pages of nineteenth-century sources. *Historical Methods: A*

- 554 *Journal of Quantitative and Interdisciplinary History* **49**, 115-131.
555 (doi:10.1080/01615440.2015.1116419).
- 556 [31] Murrieta-Flores, P., Baron, A., Gregory, I., Hardie, A. & Rayson, P. 2015 Automatically
557 Analyzing Large Texts in a GIS Environment: The Registrar General's Reports and Cholera in
558 the 19th Century. *Transactions in GIS* **19**, 296-320. (doi:10.1111/tgis.12106).
- 559 [32] Casey, A., Bennett, M., Tobin, R., Grover, C., Walker, I., Engelmann, L. & Alex, B. 2020 Plague
560 Dot Text: Text mining and annotation of outbreak reports of the Third Plague Pandemic (1894-
561 1952). *Journal of Data Mining & Digital Humanities*.
- 562 [33] Gritta, M., Pilehvar, M.T., Limsopatham, N. & Collier, N. 2018 What's missing in geographical
563 parsing? *Lang Resour Eval* **52**, 603-623. (doi:10.1007/s10579-017-9385-8).
- 564 [34] Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S. & Ball, J. 2010 Use of the
565 Edinburgh geoparser for georeferencing digitized historical collections. *Philos Trans A Math
566 Phys Eng Sci* **368**, 3875-3889. (doi:10.1098/rsta.2010.0149).
- 567 [35] van Bavel, B.J.P., Curtis, D.R., Hannaford, M.J., Moatsos, M., Roosen, J. & Soens, T. 2019
568 Climate and society in long-term perspective: Opportunities and pitfalls in the use of historical
569 datasets. *Wiley Interdiscip Rev Clim Change* **10**, e611. (doi:10.1002/wcc.611).
- 570 [36] Green, M. & Roosen, J. Biraben 2.0: A Black Death Digital Archive. Available from:
571 <https://www.radcliffe.harvard.edu/seminars-and-workshops/biraben-black-death-digital->
572 [archive](https://www.radcliffe.harvard.edu/seminars-and-workshops/biraben-black-death-digital-). [Accessed Oct 16, 2020].
- 573 [37] Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank,
574 A. & Biemann, C. 2016. A Web-based Tool for the Integrated Annotation of Semantic and
575 Syntactic Structures.
- 576 [38] Krauer, F. & Schmid, B.V. 2021 Datasets and code for "Mapping the plague through natural
577 language processing". (doi:10.5281/zenodo.4724016).
- 578 [39] Google Ireland Limited. Google Cloud Natural Language API. Available from:
579 <https://cloud.google.com/natural-language/docs/?hl=en>. [Accessed Oct 16, 2020].
- 580 [40] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. & McClosky, D. 2014. The Stanford
581 CoreNLP Natural Language Processing Toolkit.
- 582 [41] Faruqui, M. & Padó, S. 2010. Training and evaluating a German named entity recognizer with
583 semantic generalization.
- 584 [42] Explosion. spaCy v2.x. Available from: <https://spacy.io/>. [Accessed Jul 13, 2019].
- 585 [43] Explosion. 2019 de_core_news_sm-2.1.0. (
- 586 [44] Benikova, D., Yimam, S.M., Santhanam, P. & Biemann, C. 2015. GermaNER: Free Open German
587 Named Entity Recognition Tool.
- 588 [45] Geoparser Inc. geoparser.io. Available from: <https://geoparser.io/>. [Accessed Jul 8, 2019].
- 589 [46] Arnold, T. & Tilton, L. coreNLP: Wrappers Around Stanford CoreNLP Tools. Available from:
590 <https://CRAN.R-project.org/package=coreNLP>. [Accessed Oct 10, 2019].
- 591 [47] Benoit, K. & Matsuo, A. spacyr: Wrapper to the 'spaCy' 'NLP' Library. Available from:
592 <https://CRAN.R-project.org/package=spacyr>. [Accessed Oct 10, 2019].
- 593 [48] GeoNames. Available from: <http://www.geonames.org>. [Accessed Nov 11, 2019].
- 594 [49] Naber, D. Deutsches Morphologie-Lexikon. Available from: <http://danielnaber.de/morphologie/>.
595 [Accessed Apr 15, 2019].
- 596 [50] Simond PL. 1898 La propagation de la peste. *Annales de l'Institut Pasteur* **12**, 625-687.
- 597 [51] Wood, W. 1859 *The History and Antiquities of Eyam: With a Minute Account of the Great Plague
598 which Desolated that Village in the Year 1666*, Bell and Daldy.
- 599 [52] Dean, K.R., Krauer, F., Walloe, L., Lingjaerde, O.C., Bramanti, B., Stenseth, N.C. & Schmid,
600 B.V. 2018 Human ectoparasites and the spread of plague in Europe during the Second
601 Pandemic. *Proceedings of the National Academy of Sciences of the United States of America*
602 **115**, 1304-1309. (doi:10.1073/pnas.1715640115).
- 603 [53] Gritta, M., Pilehvar, M.T. & Collier, N. 2020 A pragmatic guide to geoparsing evaluation:
604 Toponyms, Named Entity Recognition and pragmatics. *Lang Resour Eval* **54**, 683-712.
605 (doi:10.1007/s10579-019-09475-3).

- 606 [54] Schmitt, X., Kubler, S., Robert, J., Papadakis, M. & LeTraon, Y. 2019. A Replicable Comparison
607 Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate.
- 608 [55] Pinto, A., Gonçalo Oliveira, H. & Oliveira Alves, A. 2016. Comparing the Performance of
609 Different NLP Toolkits in Formal and Social Media Text.
- 610 [56] Dale, R. 2018 Text analytics APIs, Part 1: The bigger players. *Natural Language Engineering*
611 **24**, 317-324. (doi:10.1017/s1351324918000013).
- 612 [57] Buntgen, U., Ginzler, C., Esper, J., Tegel, W. & McMichael, A.J. 2012 Digitizing historical
613 plague. *Clin Infect Dis* **55**, 1586-1588. (doi:10.1093/cid/cis723).
- 614 [58] Atanasiu V, Priol C, Tournieroux A & E, O. Georeferences for places of plague occurrence in
615 Europe 1347-1600. Available from: <https://bernstein.oeaw.ac.at/atlas/yersinia-description.pdf>.
616 [Accessed Mar 3, 2019].
- 617 [59] Roosen, J. & Curtis, D.R. 2018 Dangers of Noncritical Use of Historical Plague Data. *Emerging*
618 *infectious diseases* **24**, 103-110. (doi:10.3201/eid2401.170477).
- 619 [60] Carmichael, A.G. 2014 Plague persistence in Western Europe: a hypothesis. *The Medieval Globe*
620 **1**, 157-191.
- 621 [61] Benedictow, O.J. 2019 Biraben's lists of the plague epidemics of the second plague pandemic,
622 1346 - c. 1690: problems, basis, uses. *Annales de démographie historique* n°**138**, 213-223.
623 (doi:10.3917/adh.138.0213).
- 624 [62] Green, M.H. 2018 Putting Africa on the Black Death map: Narratives from genetics and history.
625 *Afriques*. (doi:10.4000/afriques.2125).
- 626 [63] Varlik, N. 2020 The plague that never left: restoring the Second Pandemic to Ottoman and
627 Turkish history in the time of COVID-19. *New Perspectives on Turkey*, 1-14.
628 (doi:10.1017/npt.2020.27).
- 629 [64] Green, M. 2014 Taking "Pandemic" Seriously: Making the Black Death Global. *Medieval Globe*
630 **1**, 27-61.

631