Supplemental material for

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*

Text S1. Re-digitization of Biraben's plague list

In 1975, the French Jean-Noël Biraben published one of the most popular treatises about plague outbreaks in Europe and the Mediterranean during the first and second pandemic [1]. With the help of dozens of other researchers, he consulted more than one thousand studies and reports and synthesized the information in two books including a large time series dataset with years and place names in the appendix [1]. Earlier historians such as Hecker [2], Martin [3] and Sticker [4] laid the groundwork with their narrative accounts of places and countries that had plague outbreaks in specific years, but it was Biraben who compiled a large amount of information in a tabular format. In the appendix of his work, he provided two lists of places or regions that reported plague in a specific year. The first table consisted of three pages addressing plague outbreaks during the First plague pandemic (between 541 – 775), followed by 73 pages addressing plague outbreaks during the Second Plague Pandemic. Both tables are grouped into different chapters according to their geographical region. The book also includes a bibliography, which the work is based on, but individual place names are not linked to a specific source. His dataset has been widely used and criticized by historians and scientists. For example Roosen and Curtis [5] mention the problem of overrepresentation of specific countries such as France or England or the underrepresentation of some areas such as the Low countries, respectively. Despite its methodological flaws, it is the best available dataset as of today for quantitative studies about the dissemination of plague in Europe during the second pandemic.

There have been two attempts to digitize and geocode the Biraben collection. The first dataset was generated by the EU-funded 'project Bernstein', which digitized historic documents and made them available in an electronic database [6]. The dataset contains 5837 observations (949 unique locations) with information about the year, the status ("region" or "location"), the modern place name, lower level administrative area names, the country name as well as the coordinates (lat/lon in WGS84). However, the original place names have not been preserved completely, which makes it difficult to compare it to other datasets or link it with the raw data. This dataset is publicly available on the project website, but it was not formally published in a peer-reviewed journal. The second dataset was digitized by Büntgen et al and was published in 2012 as a correspondence in Clinical Infectious Diseases [7]. This dataset contains 6929 observations (1171 unique locations) with information on the modern place name, the year and the coordinates. The original place names are missing as well and there is no country information. Both digitized versions are a subset of the original data set: The Büntgen data set lacks the entries for Scandinavia, the Balkans, the Beneluxe and parts of the Levante, the Bernstein dataset lacks the entries for the Maghreb, the Levante, the Southern Balkans and generally all data after the year 1600. We therefore sought to re-digitize and geocode the complete appendix four of Biraben and to provide an updated, improved digital version of Biraben's data set.

We scanned the complete appendix 4, performed optical character recognition (OCR) using Adobe Acrobat Pro and produces a plain text file. We manually corrected misaligned text and unreadable characters. The file was then read into R and a raw dataset with all observations was created. We preserved Biraben's distinction between a region and a place

(“type_orig”), whether there were doubts about a place having plague (“certain”) and in which chapter the entry occurs (“chapter”). We then batch-geocoded all locations using the methodology described in the main text. We searched only countries covered by the corresponding chapter given by Biraben, which increased the accuracy of the matches. Unclear or ambiguous locations were checked by consulting some of the original literature used by Biraben. As for the Sticker dataset, places which could not be localized exactly (including instances were Biraben described as place as “environs X”) were geocoded according to the next lower identifiable level administrative unit and marked as “approximate”. Places that could not be localized at all were marked as “unknown”. For London, Biraben occasionally lists both the city name as well as individual parishes affected by plague for a given year. This leads potentially to an inflation of data and was addressed by geocoding any parish belonging to the city of London within and without the wall (situation as of 1870) as “London”. This resulted in multiple entries for London for the same year. Researchers should apply some data cleaning or filtering before using our dataset.

Table S1. Definitions of the performance measures used in the main analysis. TP=true positives, TN=true negatives, FP=false positives, FN=false negatives.

Measure	Definition
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity (Recall, true positive rate)	$\frac{TP}{TP + FN}$
Specificity (Selectivity, true negative rate)	$\frac{TN}{TN + FP}$
Positive predictive value (PPV)	$\frac{TP}{TP + FP}$
Negative predictive value (NPV)	$\frac{TN}{TN + FN}$
F1 score	$2 * \frac{PPV * sensitivity}{PPV + sensitivity}$
Cohen's Kappa	$\frac{accuracy - p_e}{1 - p_e}$ where pe= probability of agreement by chance = $\frac{(TP + FN) * (TP + FP)}{TP + TN + FP + FN} + \frac{(TN + FN) * (FP}{TP + TN + FP}$ $TP + TN + FP + FN$

Table S2. Technical comparison of the five NLP algorithms tested. All studied tools are based on machine learning algorithms. CNN means convolutet neural network, CRF means conditional random field.

	spaCy	**Google NLP**	**Stanford CoreNLP**	**GermaNER**	**Geoparser**
model type	CNN	CNN	CRF	CRF	?
Custom model	yes	yes (Google AutoML)	yes	yes	no
open source	yes	no	yes	yes	no
pricing	free	free up to 5,000,000 characters/month	free	free	free up to 1000 API calls/month
pro-graming language	Python/Cython	?	Java	Java	?
R package	spacyr		coreNLP		
access	command line	REST API	command line	command line	REST API
supported languages	en, cn, dk, nl, en, fr, de	en, cn, dk, nl, en, fr, de, es, jp, it, korean, pt, ru	en, de, es, cn	de	?
training data set	TIGER, WikiNER	?	CoNLL 2003	GermEval 2014 NoSta-D Named Entity	?
tasks	tokenization, POS, NER	tokenization, POS, NER	tokenization, POS, NER	NER	NER LOC and geocoding
issues		POS and NER require two steps, tokenization is not identical. Umlaut must be queried as UTF-16.	Umlaut returned incorrectly with coreNLP		Discontinued as of 2020
author(s)	Explosion AI	Google Inc.	Finkel et al	Benikova et al	Geoparser Inc.
license	MIT		GNU v3	ASL 2.0	

Table S3. Comparison of identified results of tokens and entities by five different NLP and geoparser libraries.

	Gold standard	**Google NLP**	**Stanford CoreNLP**	**spaCy**	**germaNER**	**Geoparser**
Benchmark	~ 1 week	<1 min (POS) <1 min (NER)	28 mins	<1 min	11 mins	<1 min
N tokens recognised		146,340	146,743	146,767	*	*
N entities classified (%)	7884	33,925 (100)	9522 (100)	12,963 (100)	50,374	3563
Location	7884	9246 (27.25)	6989 (73.40)	10,050 (77.53)	5885 (11.68)	3563
Consumer good		765 (2.25)				
Event		2325 (6.85)				
Organization		850 (2.51)	240 (2.55)	250 (1.93)	246 (0.49)	
Other / Misc		14,875 (43.85)	817 (8.52)	1047 (8.08)	43,843 (87.03)	
Person		5755 (16.96)	1476 (15.53)	1616 (12.47)	400 (0.79)	
Work of Art		109 (0.32)				
N tokens after mapping to entities (%)						
Location	8313 (5.7)	9518 (6.5)	6987 (4.8)	10,050 (6.9)	5885 (4.0)	3659 (2.5)
Other / Not recognized	138,453 (94.3)	137,248 (93.5)	139,779 (95.2)	136,717 (93.1)	140,881 (96.0)	143,107 (97.5)

*GermaNER and Geoparser.io don't return a tokenization. Geoparser.io returns only tokens that were recognized as a toponym. GermaNER requires an a-priori tokenization of the text.

Table S4. Comparison between the Sticker and the Biraben dataset.

	Sticker	**Biraben**
N total locations	4474	11180
N exactly identified locations	4087	10644
N approximate locations	379	392
N unidentified location	8	144
N unique locations (exact & approximate)	1631	2158
Spatial coverage		
Latitude	-38°N to 75° N	12° N to 68° N
Longitude	-158 ° E to 166 ° E	-22 °E to 97° E
N countries	104	65
Types (% of total)		
Place	70.1	83.3
Administrative unit	11.4	6.7
Country	10.1	4.6
Region	7.8	4.6
Island	0.7	1
median bounding box diagonal in km (range)	17.6 (0.3 - 9007)	12.1 (0.3 - 9007)
Temporal coverage		
Years (range)	1346-1908	1346-1900
Century (% of total)		
14th	10.9	9.9
15th	6.9	15.8
16th	12	31.1
17th	23.4	30.3
18th	17.8	8.2
19th	18.8	4.7
20th	10.3	0

Figure S1. Comparison of the spatial and temporal coverage of the Sticker and Biraben dataset. (A) Relative frequency of countries (as percentage of total observations), (B) Absolute frequency of most often mentioned places, (C) yearly number of locations reported to have had a plague outbreak.

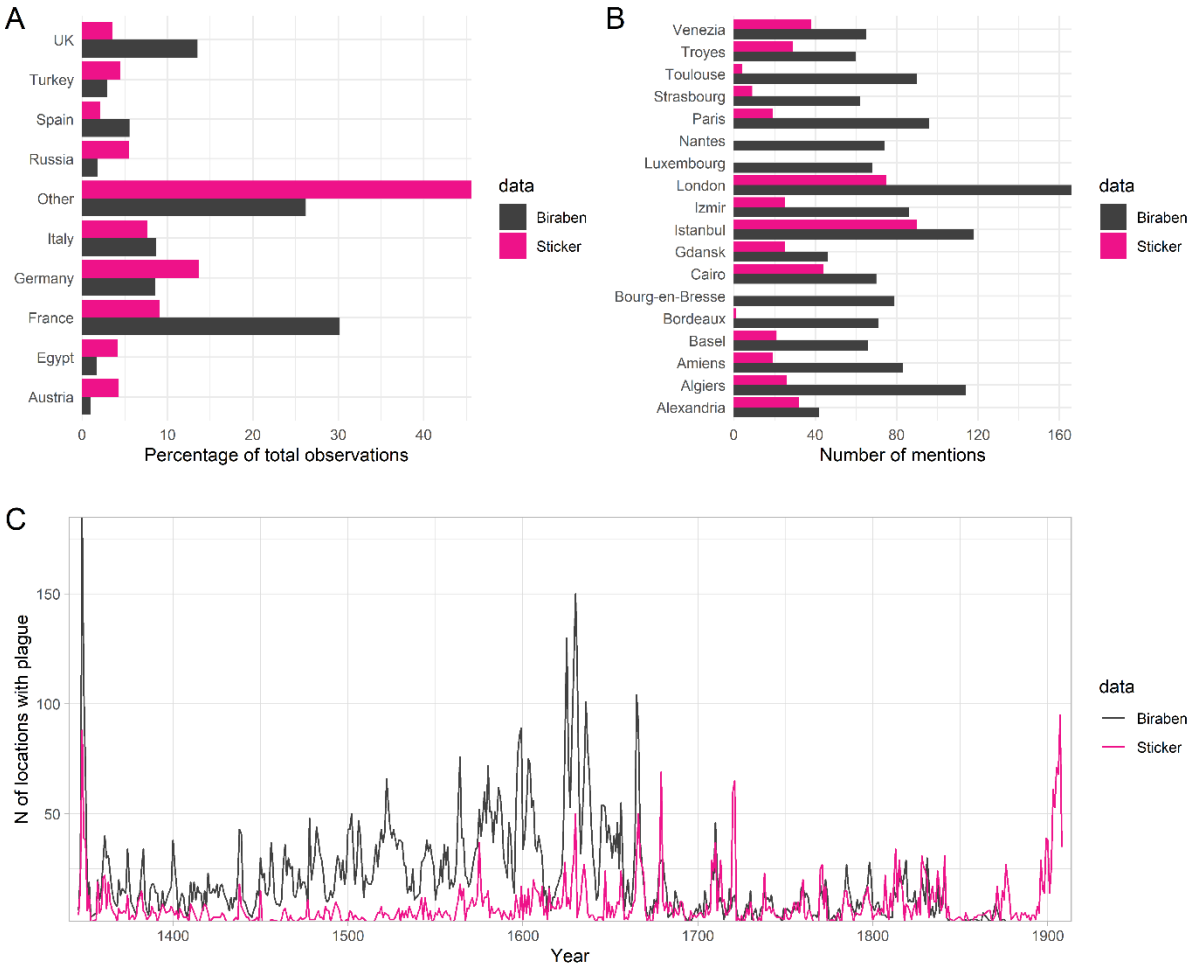
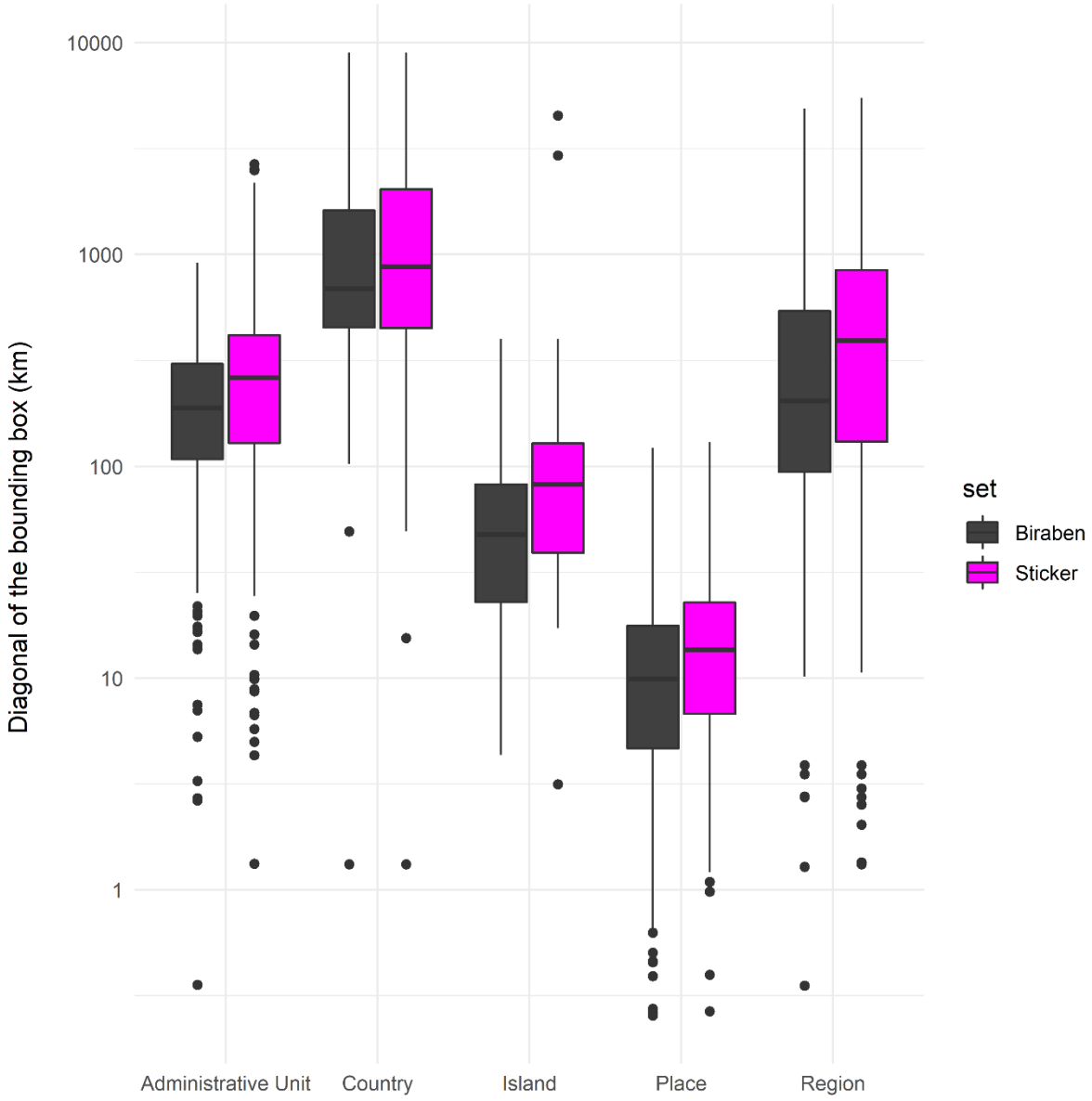


Figure S2. Diagonal of the bounding box in km according to type and dataset.



References

1. Biraben J-N. Les hommes et la peste en France et dans les pays européens et méditerranéens. Paris: Mouton; 1975.
2. Hecker JFC. Die grossen Volkskrankheiten des Mittelalters: Historisch-pathologische Untersuchungen. Gesammelt und in erweiterter Bearbeitung. Hirsch A, editor. Berlin: Enslin; 1865.
3. Martin C. Versuch einer geographischen Darstellung einiger Pestepidemien. Dr A Petermann's Mittheilungen aus Justus Perthes' Geographischer Anstalt. 1879;XXV
4. Sticker G. Abhandlungen aus der Seuchengeschichte und Seuchenlehre. Band 1: Die Pest. Giessen: A. Töpelmann; 1908.
5. Roosen J, Curtis DR. Dangers of Noncritical Use of Historical Plague Data. Emerging infectious diseases. 2018;24(1):103-10. doi:10.3201/eid2401.170477.
6. 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].
7. Buntgen U, Ginzler C, Esper J, Tegel W, McMichael AJ. Digitizing historical plague. Clin Infect Dis. 2012;55(11):1586-8. doi:10.1093/cid/cis723.