Space and time in 100 million words: health and disease in a nineteenth-century newspaper.

Catherine Porter[1], Paul Atkinson[2] and Ian Gregory[3]
[1]School of Natural and Built Environment, Queen's University Belfast, Belfast BT7 1NN
[2]Institute of Psychology Health and Society, University of Liverpool, Liverpool L69 3GB
[3]Department of History, Lancaster University, Lancaster LA1 4YT

This is an Accepted Manuscript of an article that will be published by Edinburgh University Press in the *International Journal of Humanities and Arts Computing* (2018, vol 12:2). The Version of Record will be available online at: https://www.euppublishing.com/loi/IJHAC.

Abstract

The abundance of information contained in nineteenth-century texts means the traditional 'close reading' of Victorian culture has limitations (Nicholson, 2012). With the burgeoning availability of newspapers in digital format, there is a pressing need to look at how we might effectively and efficiently use these digital resources to help answer research questions and add to key historical and geographical debates. Focusing on the analysis of a large digital corpus, this paper has two key foci: (I) to apply an innovative digital methodology, that combines corpus linguistics and geospatial technologies, to a very large corpus of newspaper texts and; (II) apply said methodology to a case study assessing the presentation of health and disease in a nineteenth-century newspaper. The paper illustrates that by linking existing techniques with new and innovative approaches it is possible to temporally and spatially analyse and map themes of interest in large digital corpora on a scale not possible through more traditional close reading methods.

Introduction

The greatest difficulty in studying the content of historic newspaper media is the sheer volume of text that must be read and processed to gain an understanding of the theme under investigation (Nicholson, 2012). Using traditional approaches, it has been difficult to fully explore these texts, but "the current generation of digital resources genuinely break new ground in what they enable us to do with nineteenth-century newspapers and periodicals and so will change the way we understand the period as a whole" (Mussel, 2012:63) This new ground speaks largely to the development of key corpus linguistic tools that facilitate a combination of close and distant reading, and the accessibility of digitised texts, tools and resources now often freely available to researchers. These resources not only allow us to study the language in historic texts but also, by combining qualitative and quantitative mechanisms, provide the tools to measure the occurrence of a theme of interest in corpora. Research by corpus linguists has varied from topics such as racism and religion, to gender (Baker, 2001; 2008, 2010; Baker et al., 2008; 2014; Bednarek, 2006; Fairclough, 1995; Teo, 2000), but hitherto there has been little analysis of health representation as depicted in early newspaper media.

The history of Victorian public health has been extensively studied and largely written in two ways: as a branch of demographic history, with the focus on statistics such as mortality rates and cause of death, and as a branch of the history of medicine (Eyler, 1979; Lambert, 1963; Szreter and Hardy, 2000; Woods, 2000; Woods and Shelton, 1997; Woods and Woodward, 1984). This research has shown little interest in the way nineteenth-century British newspapers portrayed public health and disease to the general public.[1] Research on the coverage of health and disease in nineteenth-century newspapers has so far been largely confined to non-British

[1] The exception being Porter and Porter, 1988, who discuss the use of early newspapers in publicising anti-vaccinationism and sanitation.

case studies (Almeida, 2012; Duffy, 1971; La Berge, 2002; Post, 2014), and research into health discourse in the media has tended to focus on late twentieth and twenty-first century newspaper publications (Ching, 2008; Duan, 2007; Lawrence et al., 2007). To better understand nineteenth-century media coverage of public health and disease would add to our knowledge and interpretation of the cultural history involved, from the point of view of both the production and consumption of health discourses.

Considering the former discussion on methods and theme, here, we have two main foci: (I) to robustly test a digital methodology and determine the feasibility of a semi-automated textual and geographical analysis on many millions of words and; (II) in doing so, use the case study of the presentation of public health in a nineteenth-century British newspaper and offer a preliminary examination of the health discourses it produced.

The *Era*, a London based newspaper published weekly between 1838 and 1939, has been chosen as an example of a British nineteenth-century newspaper publication for this case study (Brake and Demoor, 2009). Containing over 377 million words, it is one of a growing number of digitised nineteenth-century newspapers now available from the British Library Digital Collections, a key digital resource which currently makes accessible over 30 billion words of newspaper text. This collection has been digitised using Optical Character Recognition (OCR), a method which is known to be error prone and a great challenge to digital scholarship more generally (Tanner et al., 2009); however, tests carried out at Lancaster University showed the *Era* to be a newspaper with good quality OCR (Joulain-Jay, 2017)[2]. The *Era* was therefore chosen as a suitable newspaper for analysis in part because of its relatively strong OCR quality, but also because there is a continuous digitised series available, providing an excellent example

[2] In a separate publication we will explore the implications of OCR error on results of this type of analysis, but early indications suggest it does not seriously undermine results of analyses.

in which to trial these techniques. It should also be noted that certain corpus linguistic procedures used in this research, primarily collocation, are relatively reliable in spite of OCR errors and more so than other forms such as of keyword analysis.

The methodological and analytical approach adopted here builds on previous work conducted at Lancaster University that combines corpus linguistic analysis with Geographic Information Systems (GIS), in a suite of techniques named Geographical Text Analysis (GTA) (Gregory and Donaldson 2016; Gregory et al., 2015; Porter et al., 2015). This is the first time GTA has been used to investigate such a large corpus for a specific topic or theme. The analysis is based on three key disease groups, each of which is initially explored using corpus linguistic techniques such as frequency, collocation, and concordance analysis, to determine what the newspaper was saying about disease, and when. The temporal relationship between the frequency of disease mentions in the newspaper and deaths from these diseases is also assessed using official death statistics, and the occurrence of other terms related to public health is explored. GTA is then employed to analyse the geography of disease discussion in the newspaper at varying geographical scales. By combining these disparate techniques this paper provides, a new and robust test case for the methodologies involved, namely the sheer size and temporal coverage of the corpus being analysed, and a preliminary insight into not only how the media portrayed disease and public health in the nineteenth-century, but also where the editors focused their interests geographically.

The Disease Classifications

A list of the key nineteenth-century threats to public health was a necessary starting point. Various disease 'lists' exist, such as that composed by McKeown and Record (1962). For the purpose of this paper, the disease groups devised by Woods (2000) were chosen. Based on

three categories, ‘Crowding’, ‘Food and Water borne’, and ‘Respiration’, Woods’ scheme (Table 1) draws attention to different aspects of environmental health, respectively, housing, sanitation and air pollution. The scheme contains the most prevalent causes of death under each disease category and provides a basis for testing the digital methodology which follows.

Table 1: The three main disease categories used for analysis (Woods, 2000).

Exploring disease mentions and the relationship with death rates.

The first part of this methodology uses the corpus linguistic technique, frequency, to conduct an initial exploration of the discussion or the ‘mention’ of diseases in the *Era*. Calculating frequency using the corpus linguistic software CQPweb (Hardie, 2012) allows us to observe how frequently the keywords from each disease group were mentioned in the newspaper, analysis that would not be plausible using traditional techniques given the number of words under analysis. The raw frequency of mentions per million words was calculated per decade for each disease category (Figure 1) providing an initial picture of which diseases the *Era* was discussing and when. All three disease categories showed an initial increase from the 1840s to the 1850s followed by a decline in the 1860s. Both the Crowding and Food and Water mentions continued to follow an overall downward trend towards the end of the century. The Food and Water category had the greatest number of mentions in the early decades, but experienced a large drop in the 1860s and 1870s. In contrast to the other categories, mentions of Respiratory disease increased, tripling over the period under investigation. As the century progressed, then, the frequencies illustrate that Crowding and Food and Water diseases were discussed less while Respiratory diseases became the focus of disease mention.

Figure 1: The frequency per million words for each disease category during each decade.

To analyse the frequencies in more depth, mentions of individual diseases were also extracted. Figure 2 describes the frequency per million words for each disease that makes up the Crowding category. The *Era*'s reporting of scarlet fever, typhus, small-pox, diphtheria and measles followed a similar pattern with the highest frequency of mentions in the 1850s (whooping cough being the only exception), followed by a decline in mentions of all diseases through to the last decade of the century.

Figure 2: The frequency per million words of Crowding disease mentions, per disease, per decade.

How, then, does the frequency of newspaper comment on disease relate to levels of diseases in the population? The most suitable Victorian disease data is the reporting of death rates by cause: while some interpretation issues arise with these data, rates quoted here are reasonably reliable and have been used in previous academic publications (Hardy, 1994). Crude death rates per 1000 of the population (all ages, 1850s to 1890s: 1840s data are not available) were calculated for each disease group using the figures published by the Register-General (Figure 3) (Gregory et al., 2002).

Figure 3: Crude death rates for each disease category, per decade (source: GBHGIS).

The overall death rate from Crowding diseases exhibited a declining pattern from the 1860s onwards (Figure 3), correlating with the decline in the frequency of mentions of this disease category in the *Era*. This notable fall in death rate resulted largely from the abatement of scarlet fever, typhus and, to a lesser extent, small-pox (Hardy, 1993). The initially high death rates in

the 1850s are largely due to deaths reported as typhus,[3] and the slight rise in the 1860s was due to increased deaths from scarlet fever (Hardy, 1993). There is, then, a positive correlation between deaths from Crowding diseases and the mentions of these diseases in the *Era,* with two exceptions: diphtheria mortality was rising and measles mortality was steady, while the *Era*'s interest in both fell away. The same analysis was conducted for the Food and Water and Respiratory disease categories, an almost inverse correlation was apparent between reporting in the *Era* and trends in death rates.

What did the *Era* say about health and disease?

Frequencies are known to oversimplify the quantitative findings from texts and so they are better used in combination with other techniques such as collocation and concordance analysis (Baker, 2010). Collocation affords a statistical means of discovering which words frequently co-occur (to a statistically significant level) with words or themes of interest in a text. A concordance collects the text strings which occur around a target word in a corpus. Using both techniques in tandem divulges not only quantitative outputs of interest but also reveals the context and language of the words under analysis, in this case those that are contained within the three disease categories.

For the second step in this methodology we extracted the top twenty-five collocations with the disease keywords and uncovered numerous different words that co-occur with mentions of disease; however, analysing the collocations showed that the greatest proportion of mentions may be split into three main categories: (I) other diseases; (II) symptoms and treatment of disease, and; (III) references to deaths and illness outside Britain. Table 2 shows the top collocations for each disease category during each decade of publication.

[3] At this date some instances of typhus were misdiagnoses of typhoid

Table 2: The top collocations for each disease group in each decade.

Firstly, many of the disease keywords collocate with other diseases such as, Asiatic (cholera), 'Pleuro' and dysentery, indicative of lists of diseases in the text. Revisiting the concordances revealed that in many cases, and across all three disease categories, disease name, the corresponding number of deaths, and the place in which the deaths occurred were listed in a formal manner, *"…27 Persons died by zymotic diseases, 14 by smallpox, 11 by caleate, 70 by scarlatina, 22 by hooping-cough, 13 by croup, and 49 by tsvphusoad the various forms of fever...*" and, "*In the first nine weeks of the year (commencing 28th December , 1851), there have been registered in the Kensington district 4 deaths from scarlatina, and 17 from typhus, continued fever, and.; in Chelsea 6 from scar-latina and 7 typhus;….*" Using this form of analysis and reporting allowed us to pinpoint a recurring column in the newspaper named under various guises as, 'Health of London', 'Health of a Metropolis' and 'the Public Health', which reproduced, apparently verbatim, the weekly Registrar-General reports from the time (Figure 4). These also provide some initial insight into the geography of interest by the newspaper.

Figure 4: Excerpt from the 'Health of London' column in the *Era*.

Secondly, the collocations revealed that disease keywords often collocated with symptoms and treatment of disease; the concordances showed these were largely related to patent medicine advertisements. To consider this further, collocations that refer solely to 'treatment' of ill health (words including, cure, recipe, drink, remedy, treatment, preventative, dose, hospital, Dr. etc.) were counted and classed according to frequency, the Respiratory category being found to have the greatest number, the majority of which were again found in advertisements for medicine

such as those for the popular 'cure-all' *Chlorodyne* (Figure 5); in this excerpt the vendors refer to respiratory illnesses such as "…*coughs, asthma and bronchitis*…."

Figure 5: An excerpt from the *Era* advertising Dr J. Collis Browne's Chlorodyne (February 5, 1860).

Lastly, collocation analysis highlighted reports on deaths and illness outside of Britain, including military deaths. The related concordances reveal deaths abroad were often related to the theatre, "*MR FELIX MotRics, one of the best character actors on the American stage, died of pneumonia at his residence in New York on 13th inst. Mr Morris was born at Birkenhead, England, in 1850*." (January 27, 1900) or outbreaks of disease around the world, "*Madrid is described as a wilderness. 80,000 persons have fled from fear of the cholera, and most of those who remain barricade their doors, as was done here in the time of the Great Plague…*" (November 19, 1865). These also provide further insight into the geography of disease as published by the newspaper.

The military references noted in the collocations, words such as 'Regt.', 'Guards' and 'Brigade' were found to occur in the 1850s, relating only to the Food and Water category (deaths were largely from dysentery and diarrhoea), and specifically to the Crimean War; the newspaper publishing lists of deaths that included the name of the soldier, their regiment, and cause of death (Figure 6). Florence Nightingale was also frequently mentioned in related columns, "*Miss Nightingale and thirty-seven nurses for the sick and wounded at Scutari arrived from England in the Vectis steamer at Constantinople*" (January 7, 1855), the editors including mentions of Queen Victoria and her thoughts on the work Nightingale and her colleagues were doing in the Crimea.

Figure 6: An excerpt from the Era detailing deaths of soldiers in the Crimea (1855).

These findings warrant an exploration of the newspaper's discussion of public health more generally. The nineteenth-century evolution of public health would lead us to expect certain public health related terms to be prominent in texts from the period (Hamlin, 2015; Sturdy, 2002). This may be investigated by assessing how frequently the *Era* mentioned laws, bills and acts related to public health and disease in England and Wales (Table 3). Frequency analysis showed that terms such as 'sanitary (…*reform, commissioner, measures, authority, officers, aid, condition, arrangements, inspector, works, laws, regulations*)' and 'health of towns (…*association, commission, bill, act*)' were included in the *Era* much less frequently than the words 'health' and 'disease'. It seems the paper spent more time reporting the phenomena than the human and societal responses to them. Again, the concordances revealed the high frequency of 'health' was largely related to the aforementioned column, 'Health of London', and certain medical advertisements (the highest frequency was in the 1850s with 165.11 per million words and this decreased towards the end of the century to 41.9 per million words). The decline of both 'health' and 'disease' reflects, among other things, the *Era*'s changing editorial policy, in which the theatre and sport took an increasingly high profile (Brake and Demoor, 2009).

Table 3: The frequency per million words of public health related terminology.

The Geography of Disease.

The previously detailed portion of this methodology has shown the ability for corpus linguistics to analyse millions of words through frequency, collocation and concordance analysis and has provided detail about the *Era's* discussion of disease in the nineteenth-century. It was also

noted through the concordance analyses that place-names were commonly mentioned in relation to disease. Previous research has shown that it is valuable to explore the geography of a text using digital tools (Donaldson et al., 2017; Gregory et al., 2011; Gregory et al, 2015; Murrieta Flores et al., 2015; Porter et al., 2015) and here, in the second half of this methodology, is the first time such a large newspaper corpus has been subjected to this form of analysis. Corpus linguistics does not traditionally handle geographical exploration, but a suite of techniques developed at Lancaster University called Geographical Text Analysis (GTA) makes spatial examination of large texts possible. This is achieved using Natural Language Processing (NLP) techniques to *geo-parse* the corpus and automatically identify place-names that collocate with the disease keywords (within a span of ten words) whilst simultaneously allocating them with coordinates derived from an associated gazetteer (see Grover et al., 2010 and the Edinburgh Geoparser). At this point, a small amount of manual work is necessary to check that all place-names are allocated correctly. Errors occur, for instance, when the same place-name occurs in more than one location, or multiple spellings of a place-name exist, resulting in incorrect coordinates being assigned. During the analyses we create a 'corrections file', a list of possible errors, which we feed into the associated script to cumulatively prevent the error from repeating in future analyses.

Next, the place-names and the associated text (the concordance) derived from the GTA processing are plotted in GIS software and analysed spatially. The overall geography of disease discussion (Figure 7) shows, unsurprisingly, that the *Era*, being a London based newspaper, had the greatest number of mentions of disease in London, but also in England and other parts of Europe. There was also an apparent interest in other countries and continents such as North Africa, North America, Canada, India, China, the Philippines and Australia. These consisted

primarily of three types of report: (I) deaths from disease; (II) reporting of disease outbreaks, and; (III) diseases mentioned in advertisements for medicinal products.

Importantly, GTA preserves the concordance information related to each place-name (the text either side of the identified place-name) within the attribute table in the GIS, making it possible to investigate, in detail, what the *Era* was saying about disease and why place-names were mentioned. Outside of Europe, Egypt and Cairo appear prominently (see Figure 7) and the related reports follow the rise and fall of various disease epidemics, primarily cholera, "*The cholera at cairo is decreasing, and only forty-three deaths occurred in this town today*" (July 16, 1865); "*The PROGRESS of CHOLERA in EGYPT. That fearful scourge of the East, Cholera, has again shown its formidable head, and almost simultaneously appeared on the con-fines of Europe, Egypt, and Lesser Asia, where, since the middle of June, it has raged with even more than its customary devastation.*" (July 23, 1865). Delving deeper into the concordance reveals that a sizable percentage of the Egyptian references refer to a medicine named *'Fruit Salt'* advertised in the 1890s, Cairo being specifically mentioned in the form of a positive testimony from an individual who used it whilst living or visiting the city.

Manila was mentioned numerous times throughout the 1870s and the concordances revealed that these instances were related to a repeated mention of the treatment of cholera using the medicine *Chlorodyne,* "*Earl Russell communicated to the College of Physicians that he had received a despatch from her Majesty's Consul at Manilla, to the effect that Cholera has been raging fearflly, and that the ONLY remedy of any service was CHLORODYNE*" (April 9, 1871). References to India, including Bombay and Calcutta, do not relate to medicinal products, rather the newspaper was reporting on how disease (particularly cholera) raged throughout the century, mentioning especially the British population, particularly colonial officials and those

of the theatre fraternity, who had succumbed to the disease, "*We regret to announce the death of this talented actress, which took place at Bombay on the 2d of January, from dysentery"* and *"Major General Sir T. Valliant died of spasmodic cholera on the 22d of April, at Calcutta*"(June 22, 1845).

Figure 7: The global geography of disease (all disease categories) based on the number of mentions in the newspaper that collocate with place, illustrated using graduated symbols.

In North America the greatest number of mentions are for New York City and relate primarily to diseases such as pneumonia (Respiration) and cholera (Food and Water) in the population, chronicling disease outbreak and the deaths of specific people (again, largely individuals related to the theatre), "*A WVELL KNOWN actor, Mr. Tom Coney, who came to this country some years back with his father (the melodramatic actor and dog-trainer), died in New York on the 11th of cholera*." (September 9, 1866).

This global picture of disease interest may be refined to assess the newspaper's discussion of disease in Britain, more particularly in England and Wales (Figure 8). This cartographic depiction highlights the newspaper's focus on major urban settlements such as Newcastle Upon Tyne, Liverpool, Leeds, and of course London, but also some smaller settlements in England and Wales, such as Preston in the north, and Brighton in the south. The related concordances reveal that the places mentioned most were done so in the various contexts as described previously, examples including the formal Registrar-General reports contained in the Health of London column ("*Diarrhoea continues excessively prevalent in Norwich, Wolverhampton, Liverpool, and the Yorkshire towns*"(August 28, 1870)), registered addresses of patent-medical products such as those manufactured and sold by *Oscar Sutton and Co., Preston*, and general

death notices that mention one of the keywords from the disease categories, "*A TELEGRAM from Preston yesterday announced the death, from acute pneumonia, of Mr Joseph Pierpoint, the well-known tenor vocalist*"(June 18, 1887).

Figure 8: The geography of disease in England and Wales (all disease categories) based on the number of mentions in the *Era* and illustrated using graduated symbols.

To further investigate the projected data, GIS analysis was employed in the form of density smoothing, a form of spatial analysis that reads the point data on a map (in this case, places mentioned in the newspaper), calculates the density of said points, and produces polygons that highlight the areas of most significance in the mapped data. The analysis revealed that each disease category had the most significant number of mentions within 'London'. The related concordances showed that the largest proportion of these mentions were related to the weekly Registrar-General reports (mentioned earlier as 'Health of London' etc.) published by the newspaper, which had a London focus. An exploration of the London disease mentions reveal that most refer to an expression for the whole city such as simply 'London',[4] and therefore we should not be misled by the apparent concentration of mentions there. Numerous specific places within the city did receive mentions as well. These were mainly registration districts as was the common reporting mechanism for the Registrar-General's health reports, and include places such as Woolwich, Greenwich, Deptford, Camberwell, Whitechapel, Islington, Newington, Holborn and Chelsea, all of which lie within 2.5 miles of the River Thames (except for Norwood to the south).

[4] Note: coordinates close to the Houses of Parliament were chosen to represent the place-name 'London' and instances referring to the 'Metropolis', 'City of London' and 'London City' were also included under 'London'

The density smoothing technique was used again to refine the data further and to pinpoint the areas within London with the most significant number of mentions for each of the three disease classes (Figure 9). This revealed that for the Respiratory category (in green), mentions of disease were naming 'London' in the majority of disease discussion.[4] The Food and Water category (in pink), whilst mentioning 'London', also discussed the 'Thames', possibly reflecting mid- and late-Victorian concerns about the river as a disease vector (Horrocks, 2003). The Crowding category (in yellow), however, differs from the other two classifications in being discussed in more geographically localised terms. Several places within the city were referred to including, Marylebone (concordances reveal this was often related to the workhouse there), Westminster, Islington, Lambeth, Greenwich, Bethnal Green, Whitechapel and Poplar, all located north of the river with the exception of Greenwich, and each mentioned in relation to the number of cases of and/or deaths from specific diseases. As the concordances show, the references to Crowding diseases were being made in the Registrar-General reports and medical advertisements, and were London centric with very few cases of a local breakdown of geography noted in other towns and cities. "*In the last week, the deaths included under the zymotic class of diseases were 241; the corrected average for corresponding week being 303. There were only two deaths from small pox, and one from varicella. There were 47 from scarlatina, and 13 from diphtheria. Five from scarlatina occurred in the Kensington Town sub-district; 2 of these on the same day in one family at 8, Kensington place. Forty-four children died of measles, 3 of these in the Belgrave sub-district, 3 in the East sub-district of Islington, a in the Church sub-district of Bethnal-green (2 in one family), and 3 in Limehouse…*" (November 11, 1860).

Figure 9: Density smoothing reveals the most significant areas of disease mention for each disease category in London.

Historical Discussion

Having used disease reporting in the Era newspaper as a means of testing this combined methodology on many millions of words, numerous temporal and spatial patterns in the *Era*'s disease reporting were noted, but how should these be explained? The collocations and concordances revealed that the *Era* was printing discussion of public health and disease in various forms, primarily reproducing the Registrar-General's weekly reports, carrying patent-medical advertisements, reports on military deaths, and deaths and disease outbreaks abroad. However, the temporal frequency analysis showed that the newspaper's interest in disease was sporadic in nature (other than the weekly reports which ran up until the 1880s), with little sign that the editors had any campaigning agenda with regards to public health (specifically shown by the lack of discussion of laws, bills and public health acts related to sanitation).

The methods employed here have revealed that the *Era* was, then, making little attempt to sway Victorian opinion on issues related to public health and disease, even in London where the newspaper was based. For instance, the weekly health report which makes up the largest proportion of disease mentions up until the early 1880s (when these reports ended), and which focused primarily on London, was a column devoid of style or sensationalism; rather it was used to state the bare facts on deaths and disease in the city as reported by the Registrar-General. Other publications might have used this information to campaign for improvement in certain parts of the city. While some newspapers and periodicals had a campaigning editorial agenda about the health of the public, the *Era*'s silence in this field reminds us this was not universal (Horrocks, 2003). It chose to reproduce the weekly report for a lengthy period, which may simply have been a grateful acceptance of free editorial on the management's part. However, other copy (such as court reports) was freely or cheaply available: without access to

the publisher's records we are unable to see what motivated the choice of the weekly health report as a column-filler.

Symptoms and treatment of disease in the form of advertisements constituted proportionally the second largest mention of disease, particularly noted in the Respiratory category. Throughout the nineteenth-century the frequencies showed that the newspaper incorporated an increasing number of advertisements for medicinal products, also noted by others such as Berridge (1976). For instance, despite the *Era's* reporting on deaths and court proceedings relating to the dangers and addictive qualities of *Chlorodyne*, the newspaper increased the number of printed advertisements for the product: *Dr Collis Browne's* 'cure-all' *Chlorodyne* was first marketed in 1852, and advertisements for the product rose substantially between the 1850s and 1880s with frequencies of 1.26 and 28.42 per million words, respectively (Entract, 2015). Further work on patent medicine advertising is required, but it appears to have taken an increasing proportion of many newspapers' space, and to have focused particularly on claims to relieve respiratory disease, hence the rising level of mentions of this disease category revealed by the analysis (Figure 1).

Due to the prevalence of, and the corresponding public interest in, certain deadly diseases, one might expect to see this concern reflected in newspaper media, but this was not true for all disease categories in the *Era*. Overall, there was a tenuous link shown between the mention of diseases and deaths from disease, some disease categories fitting the pattern of death rates better than others. The Crowding disease category was the only one of the three that had a corresponding pattern of decline in mentions and crude death rate. The introduction of legislation such as sanitary reform coupled with rising living standards, changes in the virulence of some infectious agents, and increased medical understanding of disease, meant

deaths from many diseases in the nineteenth-century did decrease (McKeown and Record, 1962; Hardy, 1993). In particular, scarlet fever and typhus declined dramatically from the 1860s onwards. The newspaper may, then, have been reporting Crowding related illnesses and deaths less because there was a lower threat to the population.

The Food and Water disease group showed little correlation between deaths and mentions, with one notable exception: a correlation between one cholera outbreak in the mid-1850s, the related increase in deaths, and increased disease mentions in the newspaper. Cholera is especially interesting due to the number of people who succumbed to the disease.[5] Snow (2002) described it as a hot topic for the British media of the time as well as in other countries. Concordances in the *Era* show some evidence of this interest, the newspaper describing cholera as a "deadly and frightful" disease that raged throughout industrial England and Wales, as well as elsewhere in Europe and North America. Surprisingly, there was no obvious link between, for instance, the cholera outbreak of 1866 and reporting in the newspaper. In fact, only one 'special report' was identified from this particular outbreak titled *"*Cholera at Sea*"* (June 5, 1866). In addition to this, the concordances related to the spike in mentions of the category in the 1880s were examined but may be accounted for by the inclusion of cholera, diarrhoea and dysentery in certain 'cure-all' advertisements during that decade, rather than by disease outbreaks.

How do we account for the spatial pattern of the *Era*'s reporting of disease? Its London focus is unsurprising, but the breakdown of disease category mentions within the city is of interest. The newspaper did not distinguish between different parts of London for the Respiratory and Food and Water categories; only 'London', or 'London' and the 'Thames', respectively, were used to describe these diseases. In the Crowding category, however, the geography of mentions

[5] Cholera was an endemic disease with epidemics occurring in Britain in 1832, 1848, the early-mid 1850s, and in 1866

across the city was more fine-grained. It did not, however, concentrate on the unhealthiest places: of those it mentioned (Figure 9), some had above average infant mortality rates (a good general measure of health in a district (Williams and Mooney, 1994)) at the relevant time (for example Bethnal Green and Kensington), but the overall picture is mixed, with other places receiving a mention despite seeming to be healthier than average. It may be speculated that the reason for a more detailed spatial reporting could be the aetiology of crowding diseases, which produces outbreaks with a tighter geographical focus (Hardy, 1993).

GTA also revealed that the *Era* was providing readers with insight into death and disease globally, usually in relation to major outbreaks of disease or deaths of British subjects abroad, the various countries and continents mentioned reflecting Britain's nineteenth-century imperial position in the world as a trading nation (Figure 7). Those places highlighted by the analysis were largely those of white settlement and/or British colony, trading ports in South America and the Philippines for instance, being described in terms of the deaths or illnesses on board ships and vessels used in British trade. The geographical data reflect the flow of information from these territories towards England's capital through trading routes, but it is also indicative that the newspaper was read elsewhere; it mentions on the *Era*'s masthead that it had 'a world-wide circulation', suggesting it may have been read by (expatriate) people in other parts of the world, and therefore it was in the newspaper's interest to provide updates to readers on disease abroad.

The incidence of communicable disease, locally and globally, has long been a human concern, if only for the purpose of avoiding unhealthy places or potentially infectious travellers and their goods. Surveillance of disease events, and the drawing of attention to them when they occur, usually in a regular report, has been a frequent response (Hamlin, 2015). In Victorian England

the Registrar-General began a more sophisticated surveillance of communicable disease than had hitherto been seen, and this included a weekly report about disease in London (Higgs, 2004) which the concordance analysis reveals was republished by the *Era*. The *Era* also added more opportunistic and occasional reporting of disease, both in Britain and abroad, presumably from diverse sources including agencies. Some of these reports, such as those of disease in distant trading ports, appear to have been published to extend the newspaper's role in alerting readers to threats of disease, though this was never pursued in a systematic manner and the volume of such reports was small.

Methodological discussion and Conclusion

In recent years researchers have made huge advancements in the development of approaches for investigating digital corpora. This would not be possible without key multi and inter-disciplinary projects such as 'Spatial Humanities: Texts, GIS & Places'[6] where researchers from geography, history, computer science, English literature and digital humanities backgrounds work in tandem to produce tools and methodologies that will facilitate better research. This exploratory paper is one example of this, showing that, for the first time, we have a suite of digital techniques (GTA), that when applied have the capability to mine almost a century of information contained in one nineteenth-century newspaper corpus and produce meaningful results. In addition to this, the research shows that GTA can not only assist us in investigating change over time but also change over space. In doing so, we have actively shown that "by changing the way we interact with the contents of nineteenth-century newspapers and periodicals, we can draw attention to properties… that might otherwise be overlooked" (Mussel, 2012, p62).

[6] European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant 'Spatial Humanities: Texts, GIS, places' (agreement number 283850).

The contemporary approaches presented here, a combination of previously disparate techniques from corpus linguistics and geospatial analysis, has illustrated the possibilities these digital tools provide for unlocking key information from our past. They offer a substantial development for the digital humanities in combining close and distant reading techniques previously discussed and advocated by others (Moretti, 2013; Mussel, 2012; Nicholson, 2012), as well as contributing to the subject areas of history, geography and corpus linguistics. The scale of analysis shown here was previously inconceivable through more traditional close reading methodologies, and not only provides a new template for large corpus investigation, but represents a step-change in how we analyse and understand nineteenth-century texts through digital technologies.

In this applied case study, we not only determined how a nineteenth-century newspaper discussed disease, how this corresponded with death rates, and how this altered over time and space, but most notably we tested the feasibility of these methods on a 377 million-word corpus. Further work will expand on the methods outlined in this paper to include the investigation of other newspapers from the British Library Digital Collections and to move to the next scale, testing the feasibility of analysing many billions of words through comparative studies, the examination of different themes and language, and incorporate more detailed analysis from temporal and geographical perspectives. We also aim to expand on the historical investigation outlined in this paper by developing our arguments on the representation of health and disease in other British nineteenth-century newspapers. Of primary importance is that, moving into the future, this methodology offers a new means of investigating large digital corpora that will facilitate scholars in asking and answering key humanities questions.

Acknowledgements

A special acknowledgment goes to the staff at the British Library for their provision of, and assistance with, source material for this project.

This research has been funded by the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant 'Spatial Humanities: Texts, GIS, places' (agreement number 283850).

References

Almeida, M. (2012) "The Portuguese cholera morbus epidemic of 1853-56 as seen by the press." *Notes and records of the Royal Society,* 66: 41-53.

Baker, P., Gabrielatos, C. and McEnery, T. (2014) "Sketching Muslims: A corpus driven analysis of representations around the world 'Muslim' in the British Press 1998-2009." *Applied Linguistics* 34: 255-278.

Baker, P. (2010) *Using Corpora in Discourse Analysis*. London: Continuum International Publishing Group.

Baker, P. (2008) *Sexed texts: language, gender and sexuality*. London: Equinox.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, A.M and Wodak R. (2008) "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press." *Discourse and Society,* 19, no.3: 273-305.

Baker, P. (2001) "Moral panic and alternative identity construction in usenet." *Journal of Computer-Mediated Communication,* 7, no.1: 0.

Bednarek, M. (2006) *Evaluation in Media Discourse: Analysis of a newspaper Corpus.* London/New York: Continnuum-3PL.

La Berge, A. (2002) *Mission and Method. The Early Nineteenth-Century French Public Health Movement.* Cambridge: Cambridge History of Medicine: Cambridge University Press.

Berridge, V. (1976) "Popular Journalism and Working Class Attitudes 1854-1886: A Study of Reynolds's Newspaper, Lloyd's Weekly Newspaper and the Weekly Times" (PhD diss., University of London, Birkbeck College).

Brake, L. and Demoor, M. (2009) *Dictionary of Nineteenth-century Journalism*. London: British Library Publishing Division.

British Library Newspaper Collections, 'The British Newspaper Archive,' last modified January 1, 2016, http://www.britishnewspaperarchive.co.uk.

Ching Jen, C. (2008) SARS discourse analysis: Technoscientific race-nation-gender formations in public health discourse (PhD diss., University of Maryland).

Donaldson, C., Gregory, I.N. and Taylor, J.E. (2017) "Locating the beautiful, picturesque, sublime and majestic: Spatially analysing the application of aesthetic terminology in descriptions of the English Lake District." *Journal of Historical Geography,* 56: 43-60.

Duan, J. (2007) "The discourse of disease: the representation of SARS – the China Daily and the South China Morning Post." (MPhil diss., Lingnan University).

Duffy, J. (1971) "Social impact of disease in the late nineteenth century." *Bulletin of the New York Academy of Medicine,* 47, no.7: 797-810.

Entract, J. P. J. “Browne, John Collis (1819–1884).” J. P. J. Entract In *Oxford Dictionary of National Biography*, online ed., edited by Lawrence Goldman. Oxford: OUP, 2004. http://www.oxforddnb.com/view/article/39014 (accessed November 9, 2015).

Eyler, J., (1979) M. *Victorian social medicine: the ideas and methods of William Farr*. Baltimore: Johns Hopkins University Press.

Fairclough, N. (1995) *Media Discourse*. London: Bloomsbury Academic.

Gregory, I.N., Bennett, C., Gilham, V.L. and Southall, H. (2002) “The Great Britain Historical GIS: From Maps to Changing Human Geography.” *Cartographic Journal*, 39: 37-49.

Gregory, I.N. and Hardie, A. (2011) "Visual GISting: Bringing together corpus linguistics and Geographical Information Systems." *Literary and Linguistic Computing*, 26, no.3: 297-314

Gregory, I. and Donaldson, C. (2016) “Geographical Text Analysis: Digital cartographies of Lake District literature.” In *Literary Mapping in the Digital Age* edited by David Cooper, Christopher Donaldson and Patricia Murrieta-Flores, Routledge: Abingdon.

Gregory, I.N., Donaldson, C., Murrieta-Flores, P. and Rayson, P. (2015) “Geoparsing, GIS and textual analysis: current developments in spatial humanities research.” *International Journal of humanities and Arts Computing,* 9, no.1:1-14.

Grover, C., Tobin, R., Woollard, M., Reid, J., Dunn, S. and Ball, J. (2010) “Use of the Edinburgh geoparser for georeferencing digitized historical collections.” *Philosophical Transactions of the Royal Society, A*: 368, 3875-3889.

Hamlin, C. (2015) “The History and Development of Public Health in Developed Countries.” In *Oxford Textbook of Global Public Health.* ed by Roger Detels et al. Oxford: Oxford University Press.

Hardie, A. (2012) “CQPweb - combining power, flexibility and usability in a corpus analysis tool.” *International Journal of Corpus Linguistics*, 17: 380-409.

Hardy, A. (1993) *The epidemic streets: infectious disease and the rise of preventive medicine, 1856-1900*. Oxford: Oxford University Press.

Hardy, A. (1994) “‘Death is the Cure of All Diseases’: Using the General Register Office Cause of Death Statistics for 1837–1920.” *Social History of Medicine,* 7: 472-92.

Higgs, E. (2004) *Life, death and statistics: civil registration, censuses and the work of the General Register Office, 1836-1952*. Hatfield: Local Population Studies.

Horrocks, C. (2003) “The Personification of "Father Thames": Reconsidering the Role of the Victorian Periodical Press in the "Verbal and Visual Campaign" for Public Health Reform” *Victorian Periodicals Review*, 36: 2-19.

Joulain-Jay, A. (2017) “Corpus linguistics for history: The methodology of investigating place-name discourses in digitised nineteenth-century newspapers” (PhD diss., Lancaster University).

Lambert, R. (1963) *Sir John Simon, 1816-1904, and English Social Administration*. London: Macgibbon and Kee.

Lawrence, J., Kearns, R A., Park, J., Bryder, L. and Worth, H. (2007) "Discourse of disease: Representations of tuberculosis within New Zealand newspapers 2002-2004." *Social Science and Medicine,* 66: 727-739.

McKeown, T. and Record, R.G. (1962) "Reasons for the decline of mortality in England and wales during the nineteenth century." *Population Studies,* 16, no.2: 94-122.

Moretti, F. (2013) *Distant Reading*. London: Verso Books.

Murrieta-Flores, P., Baron, A., Gregory, I.N., Hardie, A. and Rayson, P. (2015) "Automatically analysing large texts in a GIS environment: The Registrar General's reports and cholera in the nineteenth century" *Transactions in GIS,* 19 no.2: 296-320.

Mussell, J. (2012) *The Nineteenth-Century Press in the Digital Age*. London: Palgrave Macmillan.

Nicholson, B. (2012) "Counting Culture; or, How to Read Victorian Newspapers from a Distance." *Journal of Victorian Culture,* 17, no.2: 238-236.

Porter, C., Atkinson, P. and Gregory, I.N. (2015) "Geographical Text Analysis: A new approach to understanding nineteenth-century mortality." *Health and Place,* 36: 25-34.

Porter D. and Porter R. (1988) "The Politics of prevention: Anti-vaccinationism and Public Health in Nineteenth-Century England." *Medical History,* 32, no.3: 231-252.

Post, L. (2014) "Representing disease: an analysis of breast cancer discourse in the South African Press". (MSc diss., London School of Economics and Political Science).

Szreter, S. and Hardy, A. (2000) "Urban fertility and mortality patterns." In *The Cambridge urban history of Britain*, edited by Martin Daunton, 629-672. Cambridge: Cambridge University Press.

Snow, S. (2002 "Commentary: Sutherland, Snow and water: the transmission of cholera in the nineteenth century." *The International Journal of Epidemiology,* 31, no.5: 908-911.

Sturdy, S. (2002) "Introduction: Medicine, Health, and the Public Sphere," in *Medicine, Health, and the Public Sphere in Britain, 1600-2000,* ed. Steve Sturdy, 190-204. London: Routledge.

Tanner, S., Munoz, T. and Ros, P.H. (2009) "Measuring mass text digitization quality and usefulness." *D-Lib Magazine,* 15. 7/8: 1082-9873.

Teo, P. (2000) *Racism in the news: a Critical Discourse Analysis of new reporting in to Australian newspapers. Discourse and Society.* London: SAGE Publications.

Williams, N. and Mooney, G. (1994) "Infant mortality in an ‘Age of Great Cities’: London and the English provincial cities compared, c. 1840–1910." *Continuity and Change,* 9, no.2:185-212.

Woods, R. (2000) *The Demography of Victorian England and Wales*. Cambridge: Cambridge University Press.

Woods, R. and Shelton, N. (1997) *An Atlas of Victorian Mortality*. Liverpool: Liverpool University Press.

Woods, R. and Woodward, J. (1984) eds. of *Urban Disease and Mortality: In Nineteenth-Century England*. London: Batsford Ltd.