8 Historical Data

8.1 Introduction

The previous chapter considered change over short periods of time involving health-related corpora that reflected present-day language use. There is a long-standing tradition within corpus linguistics of analysing historical language data (e.g., Taavitsainen et al., 2014). In this chapter we also consider time, yet here we consider language use over longer stretches of time using historical corpora. In dealing with data which is remote in time, more groundwork often needs to be carried out in terms of locating and cleaning texts and forming relevant research questions (as already mentioned in Chapter 3). A detailed consideration of context is also important, in order to provide interpretations and explanations of findings which are informed, for example, by historiography. Accordingly, when looking at such data, we have found that it can be very helpful to include a historian as part of the research team.

In the following sections we will look at two case studies. The first explores an issue – vaccination – that is widely discussed in public discourse, both in the past and in the present day. The second case study – on the topic of sexually transmitted diseases (STDs) – is more exploratory and faced challenges in terms of identifying relevant examples in the corpus. This led to an analysis of collocates, which in turn resulted in the identification of new research questions.

8.2 Anti-vaccination Discourse in Nineteenth-Century England

In Chapter 3 we discussed the rationale and process for the construction of the Victorian Vaccination Discourse corpus (VicVaDis) – a 3.5-million-word collection of anti-vaccination material published in England between 1854 and 1906. During that period (which roughly overlaps with the reign of Queen Victoria), a series of Vaccination Acts made the vaccine against smallpox compulsory for babies at three months of age and imposed penalties on parents who did not comply, including fines and, potentially, imprisonment. Response to the Vaccination Acts included a large-scale organised anti-vaccination movement that persisted until a Vaccination Act in 1907 made conscientious objection easier, to the extent that in practice compulsion no longer applied. Contentious as they may have been, the Vaccination Acts marked the beginning of a process which led, thanks to vaccination, to the eventual eradication of smallpox in 1980.

As noted in Chapter 3, in 1853 immunisation against smallpox in England involved the insertion of material from the pustules of people infected with cowpox – a mild disease which had been found to confer protection against the much more dangerous smallpox in humans. Previous work based on document analysis had identified three main concerns about compulsory smallpox vaccination in nineteenth-century England (Durbach, 2005; Fajri Nuwarda et al., 2022). The first concern involved potential dangers to health. This was partly because of some inherent risks involved in vaccination practices at the time, but there were also broader beliefs that vaccination interferes with what might be seen as the purity of the human body. The second concern related to the later practice of propagating the cowpox virus on the skin of calves and sheep – the resultant vaccination was viewed as the insertion of animal products into people, which raised religious objections. The third concern involved civil liberties and the mandatory nature of vaccination; critics objected to the penalties that applied to non-compliant parents, particularly from the poorer sections of society.

The availability of the VicVaDis corpus makes it possible to systematically investigate the manifestations of these different concerns in a collection of pamphlets and other texts that were widely available at the time, as well as to identify further patterns of objections. In turn, these can be compared with what is known about vaccine-hesitant views from other historical periods – particularly the present day.

In the rest of this section, we will report the main findings of an initial analysis of the VicVaDis corpus carried out by Hardaker and co-authors (2024), and complement these findings with further observations we have made by exploring the corpus, which can be freely accessed from the website of the Questioning Vaccination Discourse project (www.lancaster.ac.uk/vaccination-discourse/vicvadis/).

8.2.1 'Vaccination' and 'Compulsory' in the VicVaDis Corpus

Hardaker and co-authors (2024) approached the question of vaccination by looking at how it was represented in Victorian textual material and comparing that to how similar concepts are represented today. They began their exploration of the VicVaDis corpus by focusing on the noun *vaccination*, which is the most frequent lexical word in the corpus, with 31,734 occurrences (9,095.55 per million words). This is, of course, not surprising, given the way in which texts were selected for inclusion in the corpus in the first place. Nonetheless, starting from the most frequent lexical words provides a data-driven rationale for the exploration of the corpus. The collocation tool in the corpus search software AntConc (Anthony, 2022) was used to compute the collocates of *vaccination* within a span of five words to the left and five words to the right of the node word. They then examined the adjective *compulsory*, which is the most frequent open-class collocate of ‘vaccination’ (log likelihood = 5,583.747; effect size using mutual information = 3.032).

Out of 2,223 occurrences of the collocational pair, a random sample of 500 concordance lines were examined manually for arguments against vaccination. This revealed some patterns that were both characteristic of the historical period represented in the corpus and comparable to anti-vaccination concerns and arguments that still apply today.

In the 1800s, one of the charges that was repeatedly levelled against smallpox vaccination was that it was ineffective. In the following extract, this claim is supported by quantitative data showing that there appears to be no clear or steady decline in the number of infections over time:

> The compulsory vaccination laws came into operation in 1854 and you would naturally expect that there would be a continuous decrease in mortality. What are the facts? In 1858, 1861, 1864 and 1867, the deaths from smallpox were 6,460, 1,320, 7,684 and 2,115 respectively. (Pickering, 1871, in *Vaccination: A Letter in Reply to an Article in the “Leeds Mercury”*)

Hardaker and co-authors (2024) point out a specific parallel with contemporary discussions regarding the effectiveness of the HPV vaccine. HPV protects against a number of cancers and diseases caused by different strains of the human papillomavirus, but in particular it protects against cervical cancer and there is increasing evidence for the effectiveness of HPV vaccination campaigns in reducing cervical cancer rates (Palmer et al., 2024). There is increasing evidence for the effectiveness of HPV vaccination campaigns in reducing cervical cancer rates (Palmer et al., 2024). However, in a 2019 study, it was found that concerns about efficacy and the length of protection were among the top sources of hesitancy in relation to this vaccine (Karaphillakis et al., 2019).

Another objection in the Victorian era was that, as a human-made process, vaccination was unnatural:

> This compulsory Vaccination, which is a wanton outrage upon nature, a stupid blunder of man, betrays also an unaccountable blindness, ignorance, or entire suppression of wrong nor human slaughter, but by the application of the powers of nature to the improvement of mankind. (Halket, 1870, *Compulsory Vaccination!! A Crime against Nature!! An Outrage upon Society!!*)

Similarly, Kata (2012), in a study of anti-vaccination websites, identified among the most common ‘tropes’ the claim that vaccines should be rejected as unnatural, as compared with immunity acquired via infection or alternative approaches to prevention (see also Fasce et al., 2023).

A related argument is that disease can be prevented by living hygienically and morally, so that vaccination is unnecessary:

> The only efficient prophylactic against disease, whether smallpox, fever &c, is to be found in enlightened and faithful compliance with the laws of life and health, which these compulsory vaccination laws- by teaching people to trust in vaccination and leading them to believe that they may nourish with impunity the real causes of smallpox- set utterly and daringly at defiance. (Pickering, 1873, *The Antivaccinator and Public Health Journal*)

This is similar to an argument that is sometimes made to refuse or delay HPV vaccination for adolescent girls in particular. As HPV is submitted via sexual contact, it is sometimes (mistakenly) suggested that infection can be avoided by limiting the number of (usually women's) sexual partners. In this way, vaccination is presented as necessary only for people with supposedly 'risky' sexual habits (Hendry et al., 2013; Semino et al., 2023). More generally, the idea that 'hygienic' lifestyles are a preferable alternative to vaccination has consistently been identified in anti-vaccination arguments in studies drawing from online twenty-first-century data (Fasce et al., 2023, Kata, 2012).

Hardaker and co-authors also identify a variety of manifestations of the view that mandatory vaccinations are incompatible with civil liberties by examining a subset of concordance lines that include the word sequence *compulsory vaccination is*. Examples of what follows this word sequence include:

- 'a great infringement on that freedom which every man has a right to enjoy';
- 'a disgrace to our jurisprudence, and a shameful intrusion upon the rights of personal liberty';
- 'the largest infringement of that freedom ever yet exercised';
- 'almost as disgusting as the forced creed practised by the Inquisition in mediaeval times';
- 'therefore a tyranny that everyone should strenuously resist';
- 'a system of tyranny and torture; I use the word advisedly'; and
- 'to surrender the cardinal principle of civil and religious liberty, and to establish a precedent for the exercise of any form of tyranny' (Hardaker et al., 2024: 170).

An examination of these concordance lines also reveals how highly emotional extended metaphors were used to highlight the consequences of compulsory vaccination for citizens' freedoms:

> A wider, and deeper, and subtler *Social* Evil than universal Compulsory Vaccination is scarcely conceivable; . . . Politically, *Compulsory Vaccination* is an innermost stab of Liberty which piercing its heart, will find its courage and licaven-bom [*heaven-born*] principles and convictions in other directions an easy

> prey. State medicine can do what it Ukes [*likes*] with us, if we once let it do this. (London Society for the Abolition of Compulsory Vaccination (1879), *The Vaccination Inquirer and Health Review: The Organ of the London Society*)

Hardaker and co-authors (2024) point out that in this extract, the author of a letter to the president and members of the International Congress on Compulsory Vaccination personifies 'Liberty' as the victim of physical aggression to support the 'slippery slope' argument that accepting compulsory vaccination will inevitably lead to further and ever more pernicious erosions of personal liberties.

At the time of this writing, no vaccinations are mandatory in England and the United Kingdom more broadly. However, similar concerns about the violation of fundamental freedoms as well as the creation of inequalities were expressed during the COVID-19 pandemic at the prospect of vaccination being compulsory for people employed in certain sectors (notably health and social care) and for the purposes of international travel (Jecker, 2022).

8.2.2 Keywords in the VicVaDis Corpus

Hardaker and co-authors (2024) also utilised the AntConc software to identify the keywords – or statistically 'overused' words – in the VicVaDis corpus (see Chapter 3). This required the creation of a reference corpus from the same variety of English and historical periods. The Corpus of Late Modern English Texts v.3.1 (CLMET3.1), compiled by De Smet and colleagues (2015; http://fedora.clarin-d.uni-saarland.de/clmet/clmet.html), was identified as a suitable source of data. It contains texts produced in the period 1790 to 1920, classified by genre, for a total of 34 million words. Hardaker and colleagues extracted all texts classified under the genre labels 'treatise' and 'letters' and written between 1850 and 1907, to match as closely as possible the inclusion criteria for the VicVaDis corpus. The result was a reference corpus, VicRef, which contains 1,947,789 tokens of comparable dates and genres as the VicVaDis corpus.

Hardaker and colleagues reported the top-25 keywords (Table 8.1; reproduced from Hardaker et al., 2024: 171) and focus on four specific lexical items than can potentially refer both to what is prevented by vaccination (i.e., smallpox) and to what is allegedly caused by vaccination (i.e., a variety of side effects or vaccine harms): *death*, *deaths*, *disease*, and *diseases*. For the purposes of the present chapter, we also consider the keyword *cases*. These five keywords appear in bold in Table 8.1.

Out of 7,207 occurrences of *death/deaths*, 861 (11.9 per cent) were found to involve deaths attributed to smallpox itself (e.g., 'deaths by smallpox' or 'death after smallpox'), while 340 (4.7 per cent) were found to involve deaths attributed to vaccination (e.g., 'deaths from vaccination' and 'deaths due to vaccination'). The latter included cases where it was claimed that the

Table 8.1 *Top 25 keywords from VicVaDis when compared with VicRef, ordered by keyness (log likelihood)*

Rank	Type	Raw frequency: VicVaDis	Raw frequency: VicRef	Normalised frequency per million words: VicVaDis	Normalised frequency per million words: VicRef	Keyness (likelihood)	Keyness (effect)
1	vaccination	31,734	4	9,095.55	2.005	28,736.667	0.018
2	smallpox	21,874	4	6,269.492	2.005	19,765.969	0.012
3	vaccinated	8,876	3	2,544.025	1.504	7,988.536	0.005
4	**disease**	**8,592**	**116**	**2,462.626**	**58.142**	**6,780.952**	**0.005**
5	dr	11,186	636	3,206.114	318.781	6,459.756	0.006
6	medical	7,793	64	2,233.618	32.079	6,441.045	0.004
7	jenner	5,345	0	1,531.976	0	4,837.453	0.003
8	cowpox	4,687	0	1,343.381	0	4,241.613	0.003
9	mr	9,608	1,140	2,753.83	571.399	3,731.588	0.005
10	lymph	3,586	5	1,027.814	2.506	3,179.169	0.002
11	was	29,005	8,671	8,313.368	4,346.144	3,141.183	0.016
12	vaccine	3,517	8	1,008.037	4.01	3,085.139	0.002
13	inoculation	3,416	3	979.089	1.504	3,048.773	0.002
14	**deaths**	**3,441**	**28**	**986.254**	**14.034**	**2,844.497**	**0.002**
15	mortality	3,373	34	966.764	17.042	2,739.78	0.002
16	compulsory	2,989	14	856.703	7.017	2,554.487	0.002
17	epidemic	2,867	9	821.735	4.511	2,490.431	0.002
18	years	7,150	997	2,049.322	499.724	2,430.964	0.004
19	**diseases**	**3,059**	**40**	**876.766**	**20.049**	**2,421.166**	**0.002**
20	unvaccinated	2,184	0	625.975	0	1,975.893	0.001
21	**cases**	**6,258**	**962**	**1,793.658**	**482.181**	**1,940.183**	**0.004**
22	cannot	2,116	0	606.485	0	1,914.357	0.001
23	london	4,198	443	1,203.224	222.044	1,770.514	0.002
24	hospital	2,276	36	652.344	18.044	1,760.796	0.001
25	**death**	**3,739**	**335**	**1,071.666**	**167.911**	**1,744.884**	**0.002**

extent of deaths caused by vaccination was being covered up by not mentioning vaccination in death certificates:

> In October 1876 an official inquiry was made concerning the illnesses through vaccination of sixteen children in the Misterton district of the Gainsborough Union, of which six proved fatal, but no mention was made of vaccination in any of the **death** certificates. Of the four deaths at Norwich, the subject also of an official inquiry in 1882, only one was certified as being due to vaccination. It appeared that nine children were vaccinated in June by Dr. Guy, the public vaccinator; of these four were dead of erysipelas within three weeks of the operation, and five were seriously ill from constitutional disease. (Tebb, 1889, *What Is the Truth about Vaccination?*)

Hardaker and colleagues point out that this charge of cover-up is still made today in arguments against vaccination. For example, during the COVID-19 pandemic, it was found that many vaccine-hesitant individuals believed that the official figures for deaths caused by COVID-19 were inflated and that the actual number of deaths caused by the vaccines was concealed (Jones et al., 2023).

The keywords *disease/diseases* occur 12,078 times in the corpus and can be used to refer to smallpox itself or to diseases that are allegedly caused by the vaccine. In the former cases, occurrences of *disease/diseases* tend to be part of claims about the ineffectiveness of the vaccine:

> There appears to be no positive security against the **disease**, either by vaccination or by smallpox inoculation, and I have seen several cases where the patients have caught smallpox twice, and have each time been very severely marked; and in two instances have died of the second attack of smallpox. (Pearce, 1868, *Vaccination: Its Tested Effects on Health, Mortality, and Population – An Essay*)

As previously mentioned, such arguments are also made against current vaccines, especially in cases where vaccines offer relatively low levels of protection against disease, as in the case of vaccines against the flu and COVID-19.

In contrast, in cases such as the following extract, the noun *diseases* occurs as part of allegations that the vaccine does more harm than smallpox itself:

> For every child that dies from smallpox, forty die from **diseases** induced by Vaccination. (LSACV, 1879*, The Vaccination Inquirer and Health Review: The Organ of the London Society*)

The nineteenth-century version of the smallpox vaccine involved the injection of bodily fluids containing live cowpox virus into babies' arms. It thus carried more dangers of a variety of side effects, some of them serious. In addition, the kind of anti-vaccination material that was included in the VicVaDis corpus blamed vaccination for a very wide variety of harmful consequences.

To investigate such allegations, in this chapter we additionally consider the keyword *cases*, which has 6,258 occurrences in the corpus (1,940.183 per million words). Where the noun *cases* is not followed by references to smallpox itself (e.g., 'cases of smallpox'), it tends to be followed by references to problems, symptoms, or diseases that are presented as the consequence of vaccination. The most frequently mentioned is the skin infection erysipelas:

> The official report of the Imperial Foundling hospital at St. Petersburg for the year 1864, the cases of erysipelas after vaccination are given as 156, of whom 2 died. (Wilkinson et al., 1879, *Vaccination Tracts*)

Other references to harms presented as being caused by vaccination in the concordance lines for *cases* include *cancer, eczema, erythema, gangrenous eruption, leprosy, lupus, prurigo, septic poisoning, syphilisation, syphilitic infection*, and *tetanus*.

We do not have the space to discuss the medical and scientific basis for each of these claims. We can, however, point out that the list of harms attributed to vaccination includes a lot of variation in terms of both plausibility and seriousness. Even the modern version of the smallpox vaccine can cause or exacerbate skin problems such as eczema (Belongia and Naleway, 2003). Conversely, the list of harms also includes at least some cases that could be described in contemporary terms as mis- or disinformation. The most obvious is cancer:

> In my individual experience a number of **cases of** cancer and sarcoma seem to have been directly traceable to vaccination as a cause. (Furnival, 1906, *Professional Opinion Adverse to Vaccination*)

During the writing of this book, similarly unsubstantiated allegations were made on social media that the cancer diagnosis revealed in March 2024 by the Princess of Wales was a result of the COVID-19 vaccines.

As Hardaker and colleagues put it,

> perhaps the most intriguing aspect of VicVaDis is that it demonstrates with remarkable clarity that the modern fears around new vaccines, such as those developed in light of COVID-19, are in fact not modern at all. Each new vaccine – smallpox, HPV, MMR, COVID-19 – may have its unique components and concerns, but the genesis of the fears themselves appears to remain relatively stable over the centuries. This data suggests that we continue to struggle with how best to assess the risks posed by the diseases themselves versus those posed by the vaccines, how to protect our children when we are required to make decisions on their behalf that have potentially severe or even fatal outcomes, how to protect our rights to make those decisions in the face of contradictory advice or mandates, where and how to draw causative links in a febrile medical arena clouded with doubts and loudly competing voices, and so forth. (Hardaker et al., 2024: 172)

The availability and exploitation of historical corpora such as VicVaDis can make an important contribution to the comparative study of discourse and social phenomena across times and cultures.

8.3 Representations around Sexually Transmitted Diseases in Early English Books Online

In this section we outline a study which provides insight into how data can help us research while challenging analysts to both revise their research questions and improve the data they use. The study also shows how a shift in the contemporary context in which work is conducted may create a relevance for a research project that was not the intended goal of the project but was, nonetheless, its outcome.

The study was part of a long-standing research programme using corpus data from the early modern period to explore marginalised identities in seventeenth-century English writing. The data in question was Early English Books Online (EEBO), a digitised collection of more than 146,000 printed works in English produced largely before 1700. This database had originally been made available as scanned images which granted researchers access to these texts without the need for expensive field trips to libraries. While these scans were of limited use to corpus linguists – they could not be searched using corpus software – they did at least allow key works to be consulted at scale, with work such as McEnery (2006) using EEBO, to explore the use of 'bad language' in Early Modern English. The scans were often derived from poor-quality texts. Even so, optical character recognition (OCR) scanning of these texts made it possible for users to search for a word in order to determine which pages from which texts might be of interest. While the results were prone to a high degree of error, the search function was still helpful. The process of improving the utility of the data began in 1999 when a group called the Text Creation Partnership (TCP) was formed, consisting of University of Michigan Library, Bodleian Libraries at the University of Oxford, ProQuest, and the Council on Library and Information Resources. The TCP worked to produce high-fidelity, machine-readable, textual transcriptions of the documents in EEBO. The resulting data *was* suitable for corpus analysis. In the work reported here, EEBO version 3 was used, providing just under a billion words of data for the seventeenth century. A billion words would, in the recent past, have presented its own challenges to standard corpus search packages, exceeding by far the limits of those software packages to search the data. However, the work on EEBO reported here used CQPweb (Hardie, 2012), which allows the rapid search and analysis of multi-billion-word corpora.

While the EEBO data still presented challenges, these were challenges linguists always had to face when working with data from this period – for example, limited metadata and variant spellings which made it difficult to search for all the occurrences of a specific word. Nonetheless, EEBO TCP opened up new avenues of possibility for corpus linguists, and in the programme of work described here, a corpus linguist and a historian worked together to carry on in the spirit of the work of McEnery (2006) to look at the representation of marginalised groups such as the poor (McEnery and Baker, 2019) and prostitutes (McEnery and Baker, 2017).

In pursuing their research on marginalised groups in the early modern period, the researchers used methods from corpus linguistics; where appropriate, these were supplemented with other methods, notably geographical information systems, to explore the locations where marginalised groups resided (Baker et al., 2019). The analysis blended the close reading and critical archive research of a historian with insights provided by linguists and corpus data.

The study summarised here began with a research question in a similar vein to those discussed previously – the researchers wanted to look at the representation of people with STDs in the seventeenth century (McEnery and Baker, 2022). While EEBO addressed the issue of access to a sufficient volume of data, availability was only the first challenge. While earlier work had proved possible with the data, when the researchers sought to explore the representation of people with STDs in EEBO, they encountered a substantial problem – one of the primary terms for referring to STDs, *pox*, was also a very frequent swear word of the time. It was also a very frequent word – there were 9,960 examples of *pox* in the section of the corpus covering the seventeenth century. With so many occurrences, reading all of the concordance lines to distinguish mentions of STDs from the use of the term as a curse or an insult would have been prohibitively time-consuming. An alternative approach would be to work with a sample of the data, which would have allowed the analysts to broadly characterise the usage of the word. As a way of trying to down-sample the data in a principled way, the analysts turned to collocation (see Chapter 1).[1] This approach proved partly helpful. Some collocates clearly related to cursing, such as *rogue*, which collocated with *pox* 30 times in the corpus and never referred to STDs, being used instead as a curse (e.g., in sentences such as 'A pox on you for a rogue').[2]

However, the collocates also revealed a further problem, along with presenting two opportunities. The problem related to those examples of *pox* which clearly related to disease. As the word refers to a symptom of disease, a pustule and its consequent scarring of the skin, it is quite vague as a term. On examining the examples of *pox* clearly related to disease, the researchers could see that only in a subset of cases could they be sure of the nature of the disease causing the pock marks. Collocates such as *venerous*, occurring 12 times, always in the fixed expression *venerous pox*, clearly relate to what would now be called

[1] In the examples cited here, we used the log ratio statistic to identify collocates, using a span of five words to the left and right as our collocation window and a minimum of five co-occurrences of the word searched for and a specific collocate to filter out low-frequency cases.

[2] From *Cambridge Jests, or, Witty Alarums for Melancholy Spirit* by an anonymous author under the pseudonym 'Lover of ha, ha, he', published in 1674 by Samuel Lowndes.

syphilis. Yet when the word collocates with other STDs, such as *gonorrhoea*, it is not necessarily the case that *pox* is indicating syphilis. In some cases it probably does as it is linked to other venereal diseases (e.g., 'cures the greatest Pox, Gonorrhea's, Cankers, and all Venereal Diseases').[3] However, the word may also appear in a list of diseases and be ambiguous, as in 'the Gonorrhea, Pox, Gout, Leprosy, and other such like Diseases perfectly cured'.[4] That ambiguity is made apparent when we consider other collocates which indicate various diseases collocating with *pox* that are also linked to pock marks, notably *small* producing *small pox* (mentioned 2,001 times in the seventeenth-century data).

Yet one type of collocate always denoted an STD and was of interest from the point of view of discourse analysis – nationality terms (e.g., *French pox*). Through such examples, the agency of the infection was, in part, ascribed to foreigners. This offered an opportunity, resulting in the analysts reconsidering their research question. Rather than focusing on the construction of those afflicted with the disease, they decided to focus on the agency for the introduction of the disease. This was both tractable as a research question and somewhat contested in the literature. The insight gained from that collocation allowed the exploration of a new, and fruitful, angle on STDs in discourse in seventeenth-century England.

The second opportunity that the collocation analysis provided related to the close reading that was carried out to support the interpretation of collocates. Through looking at the examples and checking the metadata to ascertain the date of the example and the name of the text in which it was written, an impression was formed that certain collocates of *pox* increased over time, but they also increasingly appeared in certain types of texts as the century progressed. As the texts were marked with a date of publication, it was easy to explore how a term such as *French pox* was dispersed over time. There were 1,181 such examples in the seventeenth century, and when the researchers looked at the relative frequency of the term in the decades across that century, it was found that the peak of mention for the term occurred in the 1650s (1.78 examples per million words), with a later peak in the 1690s (1 example per million words). While there was no reliable metadata that enabled an examination of distribution of the terms by genre of text, scattered, compelling evidence was found that suggested a link between time and genre relating to *pox*. For example, when exploring the texts in which *pox* and *rogue* collocated, all but two were clearly plays.

To explore the link between the construction of the disease in terms of nationality on the one hand, and genre of text on the other, it was decided to

[3] From *Choice and Experimented Receipts in Physick and Chirurgery* by Sir Kenelm Digby, published in 1675 by Andrew Clark.

[4] From *The Marrow of Chymical Physick* by William Thrasher, published in 1679 by Peter Parke.

produce a genre classification of all the texts in the EEBO v.3 corpus. The process of doing that is described in McEnery and Baker (2022), but in brief the authors used the titles of the works in question, supported by a limited amount of close reading, to first develop a genre classification of the texts in the corpus and then assign texts to that classification. The classification itself was composed of five major genres (Literary, Religious, Administrative, Instructional, and Informational), which were further divided and subdivided to produce a total of 25 sub-genres and 80 sub-sub-genres. The classification was further developed and used in a subsequent project focused on the creation of a dictionary of Shakespeare's language (see Culpeper et al., 2023; Murphy, 2019). While painstaking, the genre classification task enabled the exploration of the link between nationality, STDs, and genre.

The analysis found that while there was a wide range of nationalities associated with the pox, French people were most frequently linked to it. This gave plausibility to claims in the literature that proximity was a driver in the naming of such diseases – English writers blamed a local rival power. Three main terms were used for this, in descending order of frequency: *French pox*, *French disease*, and *morbus gallicus*. By normalising the frequency of mentions to per million words, the terms were seen to peak in use in the 1650s and 1660s, fall, and then rise in frequency again in the 1690s. However, there was also an elevated mention of *Naples* with reference to the pox, as English writers did, on occasion, mention the possibility that it was in fact Naples that was the source of the disease and more specifically soldiers of the army of Charles VIII who had marched on and captured Naples. However, as Charles was a French king at the head of a French army, this also, indirectly, pinned the blame on France. So while the *Neapolitan pox* may have been linked to Naples, more broadly it was still linked to France in wider discourse.

The genre analysis of the corpus corroborated the hypothesis that the discussion of the geographical origins of STDs varied by genre and through time. The main finding of the genre analysis was that, as the century proceeded, the discussion of French pox rose substantially in one genre: medical writing. By 1690, the number of mentions of French pox (and associated variations) had grown. From the 1630s onwards, it occurred most often in medical texts, and by the 1690s, the term was barely used in any other genre. Prior to the 1630s, there was a peak of mentions of syphilis in the genre labelled as treatise (a form of formal writing systematically investigating a particular subject), in which French pox was linked to a range of topics, including anti-Catholicism and the discussion of far-off places. Peaks in this period of texts from the history genre tend to discuss the possible geographical origins of the disease.

After a series of initial difficulties working with the data, the analysts were able to both usefully refocus their initial research question and enrich the data to discover new insights into the framing of discussions of disease in terms of blame and agency. In this case, both of those converge on one party – the French. However, as the century proceeds, we see the discourse around the disease shifting markedly towards medicalisation. That medicalisation is apparent in the pattern of collocation surrounding the phrase *French pox* – if the top-10 collocates of this word are considered in the medical writing in the corpus, it is seen that they relate to other diseases (*scurvy*, *dropsy*, *leprosy*, *gout*, *diseases*), medical treatments (*dose*, *cures*, *cured*, *cure*), and the process of infection (*infected*). As the discussion of *French pox* rises in this genre, by association this medicalised framing of it also rises. Importantly, if the collocates across the whole corpus are examined, it is seen that while collocates relating to other diseases are typical of general usage, the mention of supposed cures (*cinnabar*, *guajacum*, *quick silver*), moral judgements (*evil*), and naming strategies (*called*, *Naples*) are present in general English but not medical discourse. The rise of a medicalised discourse around French pox, therefore, is also marked by a move away from a culture of blaming, moralising, and bogus cures.

At this point it appeared that the work on this data was complete. However, in the process of submitting the work for publication, an event occurred that impacted on the publication itself. The work was submitted for publication in late 2019, just before the COVID-19 pandemic struck. Early in the COVID-19 pandemic, a process began of linking the disease, and its variants, to specific geographies, just as had happened with syphilis centuries before. The initial set of reviews for the paper asked that this be acknowledged, which was done. After resubmission, the public debate around COVID-19 had moved on, and the practice of linking COVID variants to the places in which they were detected had been much reduced, to avoid stigmatising national groups. When the paper returned from its second review, the reviewer noted this and asked that the researchers remove the reference to COVID-19 from the paper. The request was discussed with the editor, and ultimately the researchers decided to leave the reference in, as the paper was precisely about how geography and disease can combine to stigmatise. COVID-19, appearing in the publication process when it did, showed just how little had changed in terms of disease-naming practices, while also being illustrative of how naming practices can be just as sensitive and harmful today as they were four centuries ago.

8.4 Conclusion

Historical data can be much more difficult to collect, particularly if one wants to ensure a fully representative corpus. Poor OCR transfers or natural spelling variation can make it especially difficult to carry out lexicographic analysis based on word frequencies in such a context. Tools like VARD (Baron and Rayson, 2008) can help introduce a degree of systematic standardisation into historical corpus data. There is also the possibility that artificial intelligence software will be able to further improve on poorly scanned historical data, though that promise lies in the future at the time of this writing.

As the two case studies have shown, one aspect of historical corpus research on health-related topics is that the findings can sometimes help shed new light on more contemporary topics, showing how discourses can sometimes remain remarkably entrenched or are able to resurface, given similar conditions. However, there is a danger when analysing historical corpora in assuming that the attitudes and language use of the past are the same as the present. Unlike the corpora examined in the previous chapter, we do not have direct experience in the time period that these corpus texts were written in, and thus we should not impose our own values on the authors of historical texts. Context is key – it is notable how the first case study in this chapter cites large chunks of texts in order to better make sense of the use of a keyword, whereas in the second case study it was necessary to classify the genres of texts before they could be analysed. Additionally, we often need to go beyond the texts themselves, to account more holistically for the time period that they were produced in, which is one of the reasons why it can be useful to work with a historian in this challenging but rewarding form of health linguistics.

References

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Tokyo: Waseda University. Available from www.laurenceanthony.net/software.

Baker, H., Gregory, I., Hartmann, D. and McEnery, T. (2019). Applying Geographical Information Systems to Researching Historical Corpora: Seventeenth Century Prostitution. In V. Wiegand and M. Mahlberg (eds.), *Corpus Linguistics, Context and Culture* (pp. 109–36). De Gruyter.

Baron, A. and Rayson, P. (2008). VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, 22 May 2008.

Belongia, E. A. and Naleway, A. L. (2003). Smallpox Vaccine: The Good, the Bad, and the Ugly. *Clinical Medicine and Research*, *1*(2), 87–92. https://doi.org/10.3121/cmr.1.2.87.

Culpeper, J., Hardie, A. and Demmen, J. (2023). *The Arden Encyclopaedia of Shakespeare's Language*. Bloomsbury.

Durbach, N. (2005). *Bodily Matters. The Anti-vaccination Movement in England, 1853–1907*. Duke University Press.

Fajri Nuwarda, R., Ramzan, I., Weekes, L. and Kayser, V. (2022). Vaccine Hesitancy: Contemporary Issues and Historical Background. *Vaccines*, *10*(10), 1595. https://doi.org/10.3390/vaccines10101595.

Fasce, A., Schmid, P., Holford, D. L., Bates, L., Gurevych, I. and Lewandowsky, S. (2023). A Taxonomy of Anti-vaccination Arguments from a Systematic Literature Review and Text Modelling. *Nature Human Behaviour*, *7*(9), 1462–80. https://doi.org/10.1038/s41562-023-01644-3.

Hardaker, C., Deignan, A., Semino, E., Coltman-Patel, T., Dance, W., Demjén, Z., Sanderson, C. and Gatherer, D. (2024). The Victorian Anti-Vaccination Discourse Corpus (VicVaDis): Construction and Exploration. *Digital Scholarship in the Humanities*, *39*, 162–74. https://doi.org/10.1093/llc/fqad075.

Hardie, A. (2012). CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics*, *17*(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har.

Hendry, M., Lewis, R., Clements, A., Damery, S. and Wilkinson C. (2013). 'HPV? Never Heard of It!': A Systematic Review of Girls' and Parents' Information Needs, Views and Preferences about Human Papillomavirus Vaccination. *Vaccine*, *25*(45), 5152–67. https://doi.org/10.1016/j.vaccine.2013.08.091.

Jecker, N. S. (2022). Vaccine Passports and Health Disparities: A Perilous Journey. *Journal of Medical Ethics*, *48*, 957–60. https://doi.org/10.1136/medethics-2021-107491.

Jones, L., Bonfield, S., Farrell, J. and Weston, D. (2023). Understanding the Public's Attitudes towards COVID-19 Vaccinations in Nottinghamshire, United Kingdon: Qualitative Social Media Analysis. *Journal of Medical Internet Research*, *25*. https://doi.org/10.2196/38404.

Karaphillakis, E., Simas, C., Jarrett, C., Verger, P., Peretti-Watel, P. and Dib, F. (2019). HPV Vaccination in a Context of Mistrust and Uncertainty: A Systematic Literature Review of Determinants of HPV Vaccine Hesitancy in Europe. *Human Vaccines Immunotherapeutics*, *15*(7–8), 1615–27. https://doi.org/10.1080/21645515.2018.1564436.

Kata, A. (2012). Anti-vaccine Activists, Web 2.0, and the Postmodern Paradigm – An Overview of Tactics and Tropes Used Online by the Anti-vaccination Movement. *Vaccine*, *30*(25), 3778–89. https://doi.org/10.1016/j.vaccine.2011.11.112.

McEnery, T. (2006). *Swearing in English: Bad Language, Purity and Power 1586 to the Present*. Routledge.

McEnery, T. and Baker, H. (2017). *Corpus Linguistics and 17th Century Prostitution*. Bloomsbury.

(2019). Language Surrounding Poverty in Early Modern England: A Corpus Based Investigation of How People Living in the Seventeenth Century Perceived the Criminalized Poor. In C. Suhr, T. Nevalainen and I. Taavitsainen (eds.), *From Data to Evidence in English Language Research* (pp. 225–57). Brill.

(2022). A Geography of Names: A Genre Analysis of Nationality-Driven Names for Venereal Disease in the Seventeenth Century. In T. Hiltunen and I. Taavitsainen (eds.), *Corpus Pragmatic Studies on the History of Medical Discourse* (pp. 23–48). John Benjamins.

Murphy, S. (2019). Shakespeare and His Contemporaries: Designing a Genre Classification Scheme for Early English Books Online 1560–1640. *ICAME Journal*, *43*(1), 59–82. https://doi.org/10.2478/icame-2019-0003.

Palmer, T. J., Kavanagh, K., Cuschieri, K., Cameron, R., Graham, C., Wilson, A. and Roy, K. (2024). Invasive Cervical Cancer Incidence Following Bivalent Human Papillomavirus Vaccination: A Population-Based Observational Study of Age at Immunization, Dose, and Deprivation. *JNCI: Journal of the National Cancer Institute*, *116*(6), 857–65. https://doi.org/10.1093/jnci/djad263.

Semino, E., Coltman-Patel, T., Dance, W., Deignan, A., Demjén, Z., Hardaker, C. and Mackey, A. (2023). Narratives, Information and Manifestations of Resistance to Persuasion in Online Discussions of HPV Vaccination. *Health Communication*, *39* (10), 2123–34. https://doi.org/10.1080/10410236.2023.2257428.

Taavitsainen, I., Hiltunen, T., Lehto, A., Marttila, V., Pahta, P., Ratia, M., Suhr, C. and Tyrkkö, J. (2014). Late Modern English Medical Texts 1700–1800: A Corpus for Analysing Eighteenth-Century Medical English. *ICAME Journal*, *38*(1), 137–53. https://doi.org/10.2478/icame-2014-0007.