1 Introduction

1.1 Why This Book?

We decided to write this book to document and share the experiences, methods, and findings associated with more than 10 years of corpus linguistic research on health communication in the Department of Linguistics and English Language and the ESRC Centre for Corpus Approaches to Social Science at Lancaster University, UK.

Lancaster University has been at the forefront of the development of corpus linguistics since the 1970s, when expertise in linguistics and computer science began to be combined to create and exploit the large electronic collections of language data known as *corpora* (singular *corpus*, from the Latin word for *body*). Alongside theoretical and empirical contributions to linguistics itself, the Lancaster corpus linguistics tradition has always been focussed on applications of corpus methods beyond linguistics and outside academia. In 2013, this concern for the potential of corpus linguistics to address questions from other disciplines and issues from society at large resulted in the foundation of the Centre for Corpus Approaches to Social Science (CASS), with funding from the Economic and Social Research Council (ESRC, part of UK Research and Innovation). In its first five years, CASS focussed on applications of corpus methods to social science disciplines such as criminology, sociology, and economics, and provided the environment in which our first health-related research blossomed, for example, with projects on metaphors in communication about cancer (Semino et al., 2018) and on patients' online feedback on the National Health Service (NHS) in England (Baker et al., 2019a). Between 2018 and 2024, further ESRC funding for CASS enabled a research programme on health communication specifically, including strands on media representations of obesity (Brookes and Baker, 2021); accounts of lived experience with anxiety, psychosis, and chronic pain (Collins and Baker, 2023; Collins et al., 2023; Semino et al., 2020); and interactions in healthcare settings (Collins et al., 2022). Further funding from ESRC and other sources enabled us to extend our work to discourses around dementia and vaccinations (Brookes, 2023; Coltman-Patel et al., 2022; Putland and Brookes, 2024), and to historical

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

issues such as representations of sexually transmitted diseases in seventeenth-century England (Baker et al., 2019b).

We carried out this work because, alongside linguists working with different methods and in many other universities in the UK and around the world, we believe (and aim to demonstrate that) linguistic expertise has a great deal to offer to research and practices around health and illness (e.g., Demjén, 2020). Much of what we do when we are ill is talk, write, and read about our symptoms, diagnoses, treatment, and outlook. Healthcare professionals communicate with patients and one another as a central part of their jobs. The experiences, causes, and consequences of illness are regular topics of discussion in mainstream media and social media. In this context, a range of linguistic methods and approaches are relevant to understanding the lived experiences of patients, carers, and health professionals; identifying problems and potential solutions in communication in healthcare settings; and investigating public representations of illness and their consequences, especially where prejudice and stigma may be involved. This requires the whole arsenal of theories and methods that linguists have at their disposal, including ethnography, conversation analysis, pragmatics, and, in our case, corpus linguistics (e.g., Brookes and Collins, 2023; Brookes and Hunt, 2021; Demjén, 2020; Hamilton and Chou, 2014; Harvey and Koteyko, 2013).

Corpus methods are particularly relevant where it is possible, necessary, and/or beneficial to collect and study health-related datasets that are too large to be analysed manually. In this book, this usually involves specialised corpora consisting of millions or tens of millions of words, although general corpora can of course be much larger than this. In some cases, large datasets were brought to us by external stakeholders, as in the case of the corpus of NHS feedback discussed in Chapters 2, 6, 11, and 12. In most other cases, we built and analysed corpora from sources where large quantities of data exist and are ethically accessible for research, such as news reports on obesity in Chapters 2, 3, and 10, and online forum posts about cancer in Chapters 5, 9, and 12. While qualitative analyses of small quantities of such data are of course extremely valuable, corpus methods make it possible to combine quantitative information about large-scale patterns with in-depth analyses of specific examples or interactions in context, as also shown in the work by our colleagues in other universities (e.g., Harvey, 2012; Kinloch and Jaworska, 2020; Mullany et al., 2015). Among other things, this in some cases can bridge an unhelpful divide in healthcare research between qualitative and quantitative methods (Greenhalgh, 2016).

Prior to writing this book, we have described the corpora, methods, and findings of our work on health communication in specialised publications (research monographs, articles in linguistic and healthcare journals, specialised edited collections) and in blog posts, podcasts, and media interviews for the general public. Through this book, we aim to present our work, and what can be

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

learnt from it, for the benefit of readers who are not already experts in corpus linguistics, healthcare research, or either. This includes students and, more generally, anyone who might seek guidance or inspiration in beginning to use corpus methods to study health communication or beginning to apply their expertise in corpus methods to health communication. To achieve this, we have not simply endeavoured to present our work in a more accessible way than in more specialised writing, but we have included details about aspects of the research that are not suitable for other forms of publication, such as different ways of formulating research questions, dealing with ethical issues, and engaging with non-academic stakeholders. In this way, we hope to make our experiences over the last 10 years as useful to our readers as we can.

It is beyond the scope of this book to provide a detailed introduction to corpus linguistics and the different tools and techniques associated with it. Several such introductions already exist, such as McEnery and Hardie (2011), O'Keeffe and McCarthy (2021), and Hunston (2022). Nonetheless, in the next section we provide a brief overview of the main corpus linguistic techniques that are referred to in the rest of the book, for the benefit of readers unfamiliar with corpus linguistics.

1.2 An Overview of Corpus Linguistic Analytical Techniques

Corpus linguistics, as we have mentioned, involves the use of tailor-made software tools to study patterns of linguistic choices in digital collections of texts, or corpora, that are too large to analyse by hand or eye alone (McEnery and Hardie, 2011). Such software tools make it possible, among other things, to study the frequencies of words, patterns of co-occurrence of words (collocations), instances of words in context (concordances), and unusually frequent words (keywords). In fact, corpus tools make it possible to carry out the same analyses at the level of grammatical categories and semantic fields. In this section, however, we will demonstrate the technique at the level of words.

More specifically, we will briefly demonstrate each of these techniques with reference to a three-million-word corpus of literature (mainly pamphlets) from the Victorian period in England that opposed vaccination against smallpox, which was made compulsory in 1853: the Victorian Anti-Vaccination Discourse corpus (VicVaDis). The composition of the corpus is described in detail in Chapter 3, and an example of exploitation of the corpus is provided in Chapter 8, based on Hardaker and colleagues (2024). The analyses carried out were obtained by loading texts into the free corpus analysis tool AntConc (Anthony, 2022; www.laurenceanthony.net/software/antconc/). Many other tools are available, and several will be mentioned in the course of this book, including CQPweb (Hardie, 2012; https://cqpweb.lancs.ac.uk),

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

Sketch Engine (Kilgarriff et al., 2014; www.sketchengine.eu), Wmatrix (Rayson, 2008; https://ucrel-wmatrix6.lancaster.ac.uk/wmatrix6.html), and WordSmith Tools (Scott, 2016; www.lexically.net/wordsmith/). Each offers a range of similar functions, along with some which are unique. It is beyond the scope of this book to provide introductions to these different corpus tools, but in all cases online guides or tutorials are available for both novice and advanced users.

1.2.1 Frequency Lists

An initial exploration of a corpus may involve the extraction of a frequency list – namely, a list of all the words included in the corpus, in decreasing order of frequency. The topmost frequent words in a corpus tend to be grammatical words, such as (in English) *the*, *of*, and *it*. To begin to explore the VicVaDis corpus, Hardaker and colleagues (2024) extract a frequency list of lexical, or open-class, words (i.e., nouns, verbs, adjectives, and adverbs). They then present the top 10 most frequent lexical words in the corpus (Table 1.1; note that in the final row, *jenner* references the surname of the doctor who is credited with the introduction of vaccination against smallpox in England). Table 1.1 includes both the raw frequencies and the relative or normalised frequencies per million words. Normalised frequencies are particularly relevant when comparing corpora of different sizes.

Hardaker and colleagues (2024) begin by studying the use of the most frequent lexical word in the VicVaDis corpus (i.e., the noun *vaccination*).

Table 1.1 *Most frequent open-class words in VicVaDis, ordered by raw frequency*

Word	Raw frequency	Normalised frequency per million words
vaccination	31,734	9,095.55
smallpox	21,874	6,269.492
dr	11,186	3,206.114
mr	9,608	2,753.83
vaccinated	8,876	2,544.025
disease	8,592	2,462.626
medical	7,793	2,233.618
years	7,150	2,049.322
cases	6,258	1,793.658
jenner	5,345	1,531.976

Adapted from Hardaker et al. (2024): 168.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

Table 1.2 *Top-10 open-class collocates of* vaccination *in VicVaDis with a –5/+5 window, ordered by log likelihood*

Collocate	Frequency	Likelihood	Effect
compulsory	2,223	5,583.747	3.032
after	1,350	1,255.124	1.639
question	1,048	1,181.757	1.839
inquirer	379	1,053.572	3.243
anti	769	991.075	1.994
acts	453	947.557	2.697
against	1,055	846.510	1.504
act	694	749.524	1.793
league	251	532.783	2.723
tracts	184	474.361	3.087

Adapted from Hardaker et al. (2024): 169.

1.2.2 Collocates

One way of understanding how particular words are used in a corpus is to look at what other words tend to occur around them more frequently than one would expect by chance. Such words are known as the *collocates* of that particular word of interest, or *node* word. Patterns of collocation can reveal the meanings and associations of words in a particular corpus.

Hardaker and colleagues (2024) extract from the VicVaDis corpus the collocates of the noun *vaccination* – the most frequent open-class word. Table 1.2 provides the top 10 open-class collocates of *vaccination*. They were identified within a window of five words to the left and five words to the right of the node word, and by means of two statistical measures: a measure of statistical significance, log likelihood, which captures the probability that the relationship between two words may occur by chance (see the 'Likelihood' column); and a measure of effect size, which captures the strength of the collocation between the node word and each collocate. The table also provides the overall number of occurrences of the collocational pair in the corpus (see the 'Frequency' column).

Hardaker and colleagues (2024) then focus on the collocation between *compulsory* and *vaccination*, to identify objections to the mandatory nature of the smallpox vaccine in the VicVaDis corpus.

1.2.3 Concordances

A more detailed, qualitative way of studying the use of particular words or combinations of words in a corpus is to obtain a *concordance* (i.e., all instances of that word or combination of words in context). Figure 1.1

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

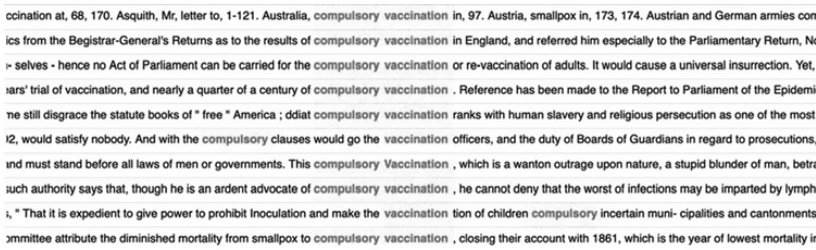


Figure 1.1 Extract from the concordance of the collocational pair *vaccination* and *compulsory*.

provides an extract from the concordance of *compulsory* collocating with *vaccination* in the VicVaDis corpus.

Hardaker and colleagues (2024) explore the concordance to identify the main objections to compulsory vaccination in the corpus, such as, for example, that it is unnatural and a violation of civil liberties.

1.2.4 Keywords

Finally, *keyness* analysis makes it possible to identify the distinctive vocabulary in a corpus of interest (the 'target' corpus) by comparing the relative or normalised frequencies of words against a corpus that can be seen as a relevant norm (the 'reference' corpus). The resulting 'keywords' are words that are statistically 'overused' in the target corpus as compared with the reference corpus.

Hardaker and colleagues (2024) extract the keywords in the VicVaDis corpus by comparing it against a reference corpus labelled the VicRef corpus (see Chapter 8 for more detail). The reference corpus consists of a tailor-made nineteenth-century corpus containing texts from similar genres but involving a wide variety of topics. The top 25 keywords are presented in Table 1.3. The table includes raw and normalised frequencies for both corpora, as well as scores for the same two statistical measures we have mentioned in the section on collocations.

In corpus analyses, keywords are often grouped according to particular themes, based on semantic similarity, grammatical similarity, or a combination of the two. Subsequently, selected groups are subjected to more detailed analysis. For example, as we show in Chapter 8, Hardaker and colleagues (2024) look at concordance lines for the keywords *death*, *deaths*, *disease*, and *diseases* as a grouping of semantically related terms in order to identify the different kinds of harms that are attributed to vaccination, and the ways in which those harms are presented.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

Table 1.3 *Top-25 keywords from VicVaDis when compared with VicRef, ordered by keyness (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

1.3 The Structure of This Book

In this book we demonstrate the kinds of questions, settings, and datasets that can be involved in corpus-based studies of health communication, and the variety of tools that can be employed to carry out these studies. We aim to do this in a way that is maximally useful to readers who are not already experts in this area, and especially students. Thus, we have structured the chapters according to the likely sequential stages of the research process: from research questions to dissemination, with some chapters on selected topics for analysis in the middle. Each chapter demonstrates the relevant stage of research or topic with reference to two or three specific projects that at least one of us has been involved with at Lancaster University. While all five co-authors take responsibility for the whole book, in this section we indicate who is particularly responsible for each chapter. In the chapter itself, we also mention team members not involved in this book, where relevant.

Following the present introductory chapter (by Semino), Chapter 2 (by Baker and Semino) is devoted to the formulation of research questions. We discuss the different processes through which research questions can be identified and developed in corpus-based research on health communication, depending on the nature of the project and the degree of involvement of different stakeholders. Three studies are considered, in order to compare how various research questions were formulated. The first study involved the analysis of press representations of obesity (Brookes and Baker, 2021). In this study, the researchers developed their own research questions in a variety of ways, including, for example, by drawing from the non-linguistic literature on obesity. The second study focussed on the McGill Pain Questionnaire (MPQ) – a well-known language-based diagnostic tool for pain. A pain consultant asked the researchers if they could help understand why some patients find it difficult to respond to some sections of the questionnaire. In response, the researchers formulated a series of questions that could be answered using corpus linguistic tools and identified some issues with the questionnaire that address the pain consultant's concerns (Semino et al., 2020). The third study (Baker et al., 2019a) involved the analysis of patient feedback on the NHS. The researchers were approached by the NHS Feedback Team and given 12 questions that they were commissioned to answer by means of corpus linguistic methods.

In Chapter 3 (by Brookes, Collins, and Semino), we reflect on different approaches to data collection, drawing on our own experiences of corpus creation to highlight the opportunities and challenges associated with building corpora from health communication data. We begin with a case study based on a purpose-built corpus of news articles on the topic of obesity, collected from the LexisNexis online news repository. We focus on theoretical considerations attending to corpus design (i.e., the 'aboutness' of the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

texts collected and the balance and representativeness of the corpus as a whole), as well as practical challenges involved in processing texts provided by repositories such as LexisNexis to make them amenable to corpus analysis (e.g., removing repeated texts, noise, boilerplate text, etc.). The second case study focusses on how corpus linguists might work with existing datasets of health communication data – in this case, transcripts collected by research collaborators conducting ethnographic research in the context of Australian emergency departments. We discuss the ways in which data collected by researchers for the purposes of different kinds of analysis is likely to require some pre-processing before we can consider it a corpus suitable for corpus-based analysis. The third case study is concerned with the creation of the VicVaDis corpus, which we mentioned earlier. We discuss the challenges and decisions involved in sourcing historical material from existing databases, selecting a principled set of potential candidate texts for inclusion and using optical character recognition (OCR) software to convert the texts into a format that is appropriate for the use of corpus tools.

In Chapter 4 (by Semino and Brookes), we consider ethical issues in healthcare communication research through two case studies. The first case study looks at a relatively straightforward situation involving a study of the Pain Concern online forum. Data from the forum was provided by HealthUnlocked, a company that runs a large number of online communities related to health. One advantage of using their service was that HealthUnlocked took care of relevant legal requirements concerning ethics and only shared data from contributors to the forum who had agreed for their posts to be used for research purposes. The second case study relates to the study of dementia and brings into focus the difficulties of working with multiple datasets and a range of stakeholders. The data collection for this project involved public health communication in terms of news media and external communications from support services, including social media. As such, it presents scenarios that are common to studies of health communication and thereby offers instruction in how to navigate related ethical concerns.

In Chapter 5 (by Collins and Semino), we are concerned with documenting and investigating sequential aspects of health-oriented interactions and the particular challenges this poses for corpus-based research. We describe two case studies to demonstrate how conventional corpus procedures can be augmented with other linguistic approaches to facilitate a critical examination of the meaningful relationships between parts of the data that might otherwise be separated in corpus analysis, as individual texts or as participant turns in a discussion, for example. The first case study involves an approach that was developed through an investigation of the Spoken British National Corpus (BNC) 2014 to examine interactional language texts in terms of functional discourse units (Biber et al., 2021; Egbert et al., 2021). This coding framework

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

is applied to a sample of anxiety support forum data in order to document, quantify, and evaluate how various communicative purposes are formulated in forum posts and are met with different types of response. The second case study is an investigation of a specific discussion thread from an online forum dedicated to cancer – one that is explicitly dedicated to irreverent verbal play about the illness. We show how a corpus approach enabled the identification of relevant humorous metaphors and made it possible to identify recurrent lexical and grammatical features that serve important functions for facilitating discussion around sensitive topics, maintaining a coherent identity and contributing to a sense of community.

In Chapter 6 (by Brookes), we turn to how it is possible to use demographic metadata to study identities in health-related corpora. We employ two case studies to demonstrate and compare two broad approaches to identifying and studying demographic characteristics, based on research on patient feedback on NHS services in England. The first case study compares how patients of different age and sex groups evaluate healthcare services and, more specifically, how they use distinct linguistic and rhetorical strategies to do this (see Baker et al., 2019a). The corpus was encoded with demographic metadata which allowed the researchers to explore the language used by people of different age and sex identity groups when evaluating the care and treatment they received for cancer (see Brookes and Baker, 2022). For the second study, a different corpus of more general patient feedback was used, one which did not contain demographic information metadata about patients' identities. Instead, targeted searches were used to identify patients' demographic characteristics based on cases where they made those characteristics explicit within their feedback (e.g., through statements like 'I am a 55-year-old woman'). In contrasting these case studies, we also evaluate the two different approaches taken, considering the affordances and limitations of both. Taken together, the case studies demonstrate how language and identity can be explored in corpora with and without reliable demographic metadata.

In Chapter 7 (by Baker), we consider how language changes over short time spans can be examined using corpus-assisted methods. We present three case studies which consider time in different ways. The first involves the corpus of patient feedback relating to cancer care, as described in Chapter 6. This data had been collected for four consecutive years, so to compare change over time, a technique called the coefficient of variation was used to identify lexical items that had increased or decreased over time. These items were examined through concordance lines in order to uncover some of the strongest trends in terms of patient satisfaction. The second case study considered UK newspaper articles about obesity, ranging from 2008 to 2017. To examine changing themes over time, a combination of keyness and concordance analyses was employed in order to identify which themes in the corpus were becoming more or less popular over

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

time. We show how references to individual causes of obesity (e.g., diet or biological determinants) had become more popular over the years, whereas societal causes (e.g., the role of education, government, advertising, or businesses) had decreased. Additionally, the analysis considered time in a different way, by using the concept of the annual news cycle. To this end, the corpus was divided into 12 parts, consisting of articles published according to a particular month, and the same type of analysis was applied to each part. Annual patterns were found around the reporting of obesity (e.g., readers were advised to join a gym in January, whereas sleep and yoga were suggested as weight loss options in February). The third case study involves an analysis of a corpus of forum posts about anxiety from 2012 to 2020. In this study, time was considered in terms of the age of the poster. Younger posters tended to use more catastrophising language and help-seeking posts, whereas older posters tended to offer different forms of advice, which became increasingly less focussed around medical intervention, the older they were. However, time was also considered in terms of the amount of involvement that a poster had with the forum. It was found that posters progressed from initially seeking advice and providing their personal histories to increasingly taking on an emotionally supportive or advice-giving role. The more experienced posters characterised their relationship with anxiety as a learning experience or journey.

In Chapter 8 (by McEnery and Semino), we look are combined at the use of historical corpora in the study of language relating to health. We present two case studies – one where an issue is well understood and discussed publicly, the other where there was a clear issue with the framing of a discussion. The first case study explores the VicVaDis corpus, briefly introduced earlier in this chapter. Different corpus techniques to show the main anti-vaccination arguments in the corpus and to point out parallels with present-day anti-vaccination discourse. The second case study looks at the emergence of venereal disease in the seventeenth century using the Early English Books Online corpus. We show how, by examining the collocates of the word *pox*, it is possible to identify relevant uses of the word (e.g., those which referred to venereal disease as opposed to those which do not). Additionally, we show that through the investigation of one type of collocate (words referring to geographical locations), the analysis was taken in an unexpected but rewarding direction.

Chapter 9 (by Baker and Semino) considers how the experience of illness is represented linguistically, focussing on two contexts. In the first case study, collocational patterns were examined in order to show how people represented the word *anxiety*. Different patterns around anxiety were grouped together in order to identify oppositional pairs of representation (e.g., medicalising/normalising). The second case study involved an examination of the ways in which cancer was constructed in a corpus of interviews with and online forum posts by people with cancer, family carers, and healthcare professionals. Using

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

a combination of manual analysis and corpus searches, it was possible to consider how metaphors were used to convey a sense of empowerment or disempowerment in the experience of cancer. More specifically, the analysis of metaphors around cancer revealed insights into people's identity construction and the relationships between doctors and patients.

In Chapter 10 (by Baker and Collins), we demonstrate how corpus approaches support the study of various social actors, which in the healthcare context can include healthcare professionals, patients, caregivers, and even manifestations of illness. Our first case study investigates how representations of people with obesity in the UK press contribute to stigmatisation. The analysis orients around the naming strategies to collectively and individually refer to people with obesity, as well as the adjectives used to describe them and the activities that they are reported to be involved in. For example, we demonstrate a high degree of shaming in the UK press using informal, dehumanising labels such as 'fatties', 'lardy', and 'blob'. Furthermore, we show that people with obesity are regularly held up as figures of ridicule and obesity is discussed in the context of social deviance, foregrounded when reporting on perpetrators of crimes. In the second case study, we similarly discuss referential strategies, descriptions of traits, and the capacity to carry out different kinds of actions in the context of voice-hearing, to critically consider the different degrees to which people who experience psychosis personify their voices. To facilitate this analysis, it was necessary to develop a means of corpus annotation and adapt procedures for quantifying linguistic features that we argue can be more generally applied to investigations of social actors. We discuss how this corpus approach enables researchers and other stakeholders to track these representations in the reports of those with lived experience over time and consider the implications of a social actor model for therapeutic interventions to support those with chronic mental health issues.

Chapter 11 (by Brookes, Collins, and Semino) introduces the concept of legitimation in discourse and considers how it might function and be studied in the context of health(care) communication. First, we look at how contributors to the online parenting forum Mumsnet use labels denoting attitudes towards vaccinations, such as 'pro-vax' and 'anti-vax'. We point out how labels that involve opposition to vaccinations, such as 'anti-vaxxer', tend to collocate with negation, and then consider in detail how people justify negating the applicability of the label to themselves. This reveals a range of different concerns around vaccinations. We then draw on a case study of patient feedback (see also Chapter 6) which examined how patients legitimate their perspectives and the evaluations they gave in their feedback. For example, this included patients representing themselves as experienced users of healthcare services (or 'expert patients'). Additionally, some patients used aspects of their identities in order to position themselves as requiring attention, while others engaged in linguistic

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

techniques such as second-person pronouns to imply that their experiences could be generalised to other patients. Overall, the chapter underscores the need for close, qualitative examination of words and wider linguistic devices within their broader textual and health(care) contexts in order to interpret the legitimatory functions of given linguistic patterns.

Chapter 12 (by Baker and Semino) discusses the potential opportunities and challenges associated with disseminating the findings of corpus-based approaches to health communication, which also apply more generally to interdisciplinary research and collaborations between researchers and non-academic stakeholders. The first case study involves work on patient feedback with members of the NHS who had provided a list of questions for us to work on. We discuss the importance of and challenges around building and maintaining relationships with members of this large, changing organisation, while also outlining how we approached the dissemination of findings, both in academic and non-academic senses, and the extent of our impact. The second case study considers the experience of disseminating findings from the project on metaphors and cancer introduced in Chapter 5, focussing particularly on writing for a healthcare journal, dealing with the media, and going beyond corpus data to create a metaphor-based resource for communication about cancer.

Chapter 13 (by Baker) concludes the book by presenting a synthesis of the previous chapters, beginning by asking the question, 'What have our experiences taught us about health communication that we didn't know?' We go on to examine lessons we learnt about carrying out corpus-based research on health communication, offering practical advice and tips for people who might be carrying out similar kinds of studies. We then consider the limitations of a corpus-based approach and end by looking to the future: what changes have taken place since we completed our analyses? What kinds of developments in the field of healthcare and in corpus linguistic analysis have occurred recently? And what avenues of research into health care do we believe are potentially interesting to investigate next?

We hope that this book will enable and inspire readers to pursue their own investigations, going beyond what we and others have achieved so far.

References

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

Baker, P., Brookes, G. and Evans, C. (2019a). *The Language of Patient Feedback: A Corpus Linguistic Study of Online Health Communication*. Routledge.

Baker, H., Gregory, I., Hartmann, D. and McEnery, T. (2019b). Applying Geographical Information Systems to Researching Historical Corpora: Seventeenth Century

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

Prostitution. In V. Wiegand and M. Mahlberg (eds.), *Corpus Linguistics, Context and Culture* (pp. 109–36). De Gruyter.

Biber, D., Egbert, J., Keller, D. and Wizner, S. (2021). Towards a Taxonomy of Conversational Discourse Types: An Empirical Corpus-Based Analysis. *Journal of Pragmatics*, *171*, 20–35. https://doi.org/10.1016/j.pragma.2020.09.018.

Brookes, G. (2023). Killer, Thief or Companion? A Corpus-Based Study of Dementia Metaphors in UK Tabloids. *Metaphor and Symbol*, *38*(3), 213–30. https://doi.org/10.1080/10926488.2022.2142472.

Brookes, G. and Baker, P. (2021). *Obesity in the News: Language and Representation in the British Press*. Cambridge University Press.

Brookes, G. and Collins, L. (2023). *Corpus Linguistics for Health Communication: A Guide for Research*. Routledge.

Brookes, G. and Hunt, D. (2021). *Analysing Health Communication: Discourse Approaches*. Palgrave Macmillan.

Collins, L. C. and Baker, P. (2023) *Language, Discourse and Anxiety*. Cambridge University Press.

Collins, L. C., Brezina, V., Demjén, Z., Semino, E. and Woods, A. (2023). Corpus Linguistics and Clinical Psychology: Investigating Personification in First-Person Accounts of Voice-Hearing. *International Journal of Corpus Linguistics*, *28*(1), 28–59. https://doi.org/10.1075/ijcl.21019.col.

Collins, L. C., Gablasova, D. and Pill, J. (2022). 'Doing Questioning' in the Emergency Department (ED). *Health Communication*. Online first. 1–9. https://doi.org/10.1080/10410236.2022.2111630.

Coltman-Patel, T., Dance, W., Demjén, Z., Gatherer, D., Hardaker, C. and Semino, E. (2022). Am I Being Unreasonable to Vaccinate My Kids against My Ex's Wishes?' – A Corpus Linguistic Exploration of Conflict in Vaccination Discussions on Mumsnet Talk's AIBU Forum. *Discourse, Context & Media*, *48*, 100624. https://doi.org/10.1016/j.dcm.2022.100624.

Demjén, Z. (ed.) (2020). *Applying Linguistics in Illness and Healthcare Contexts*. Bloomsbury.

Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T. and Baker, P. (2021). Identifying and Describing Functional Discourse Units in the BNC Spoken 2014. *Text & Talk*, *41*(5–6), 715–37. https://doi.org/10.1515/text-2020-0053.

Greenhalgh T. (2016). *Cultural Contexts of Health: The Use of Narrative Research in the Health Sector*. Copenhagen: WHO Regional Office for Europe. https://iris.who.int/handle/10665/326310.

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.

Kinloch, K. and Jaworska, S. (2020). Using a Comparative Corpus-Assisted Approach to Study Health and Illness Discourses across Domains: The Case of Postnatal Depression (PND) in Lay, Medical and Media Texts. In Z. Demjén (ed.), *Applying Linguistics in Illness and Healthcare Contexts* (pp. 73–98). Bloomsbury.

Hamilton, H. and Chou, W. S. (eds.) (2014). *The Routledge Handbook of Language and Health Communication*. Routledge.

Hardaker, C., Deignan, A., Semino, E., Coltman-Patel, T., Dance, W., Demjén, Z., Sanderson, C. and Gatherer, D. (2024). The Victorian Anti-Vaccination Discourse

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001

Corpus (VicVaDis): Construction and Exploration. *Digital Scholarship in the Humanities*, *39*, 162–74. https://doi.org/10.1093/llc/fqad075.

Harvey, K. (2012). Disclosures of Depression: Using Corpus Linguistics Methods to Examine Young People's Online Health Concerns. *International Journal of Corpus Linguistics*, *17*(3), 349–79. https://doi.org/10.1075/ijcl.17.3.03har.

Harvey, K. and Koteyko, N. (2013). *Exploring Health Communication: Language in Action*. Routledge.

Hunston, S. (2022). *Corpora in Applied Linguistics*, 2nd ed. Cambridge University Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014). The Sketch Engine: Ten Years On. *Lexicography*, *1*, 7–36. https://doi.org/10.1007/s40607-014-0009-9.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Mullany, L., Smith, C., Harvey, K. and Adolphs, S. (2015). 'Am I Anorexic?' Weight, Eating and Discourses of the Body in Online Adolescent Health Communication. *Communication & Medicine*, *12*(2–3), 211–23. https://doi.org/10.1558/cam.16692.

O'Keeffe, A. and McCarthy, M. J. (2021). *The Routledge Handbook of Corpus Linguistics*, 2nd ed. Routledge.

Putland, E. and Brookes, G. (2024) Dementia Stigma: Representation and Language Use. *Journal of Language and Aging Research*, *2*(1), 5–46. https://doi.org/10.15460/jlar.2024.2.1.1266.

Rayson, P. (2008). From Key Words to Key Semantic Domains. *International Journal of Corpus Linguistics*, *13*(4), 519–49. https://doi.org/10.1075/ijcl.13.4.06ray.

Scott, M. (2016). *WordSmith Tools* (Version 7). Lexical Analysis Software.

Semino, E., Demjén, Z., Hardie, A., Payne, S. and Rayson, P. (2018). *Metaphor, Cancer and the End of Life: A Corpus-Based Study*. Routledge.

Semino, E., Hardie, A. and Zakzrewska, J. M. (2020). Applying Corpus Linguistics to a Diagnostic Tool for Pain. In Z. Demjén (ed.), *Applying Linguistics in Illness and Healthcare Contexts* (pp. 99–128). Bloomsbury.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 10:31:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.001