3 Collecting Data

3.1 Introduction

In this chapter we reflect on different approaches to identifying the data that we analyse using corpus methods, since data selection has significant implications for the kinds of observations that we can make about health communication. (For a general introduction to corpus construction, see Reppen (2022).) As a fundamental part of the research process, it pays to approach data collection early on. That being said, our research aims, what we know from our reading of the relevant literature and what types of health communication are available for study can all inform how we identify, collect and record data. We report on three different approaches to data collection that utilise language content that is variously 'ready-made' for corpus analysis; in doing so, we demonstrate how the construction of corpora is often guided by research imperatives, practical and ethical concerns (see also Chapter 4), and data formatting.

The aspects of data collection discussed in this chapter will be pertinent to any kind of corpus construction, though there are particular concerns relating to the highly sensitive and personal nature of health communication data. We must also consider the extent to which the presence of a researcher in naturally occurring health communication contexts might disrupt and therefore have unfavourable impacts on the delivery of healthcare. Nevertheless, documenting healthcare interactions in their most naturally occurring state, along with relevant contextual metadata, is crucial for generating insights that can contribute to the optimisation of care.

Although we can refer to large, general corpora to investigate how, for example, people discuss health topics in general conversation, the corpora used in health-related research are most likely to be specialised. It is often the case that such corpora are purpose-built for the study, requiring some thought in terms of their design and construction. Targeting specific kinds of health content can mitigate the extent to which the corpus is representative of the perspectives of a general population. On the other hand, the population might also be targeted for its specialisation; for instance, we might want to focus on the contributions of those who have lived experience of the health topic.

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

Specialisation may also refer to a specific time period, though with historical data, we may be further restricted simply to what is available – a concern that is not exclusive to health topics.

The case studies discussed in this chapter focus on the preparation of data for the purposes of corpus-assisted research. In other words, they reflect the efforts of the researchers in facilitating analytical procedures germane to corpus analysis. To this end, we consider questions relating to what texts should be included in the corpus, how much data is required and how the files in the corpus should be organised. We reflect on the impacts of preparing those files in a machine-readable format, accounting for our decision-making on what features of the texts to include and whether we apply any kind of annotation. Finally, we discuss the importance of associated metadata and how this is incorporated into the corpus design, to enrich our interpretation of the corpus outputs.

In reporting our different approaches to data collection and corpus construction, we set out to inform readers of the key moments when it is beneficial to reflect on the purposes of the research and offer a view of some of the options that can help make this process more efficient and productive for those research purposes. We begin by looking at the fundamentals of creating a new dataset to investigate representations of obesity.

3.2 Compiling a News Corpus about Obesity

The first corpus we discuss in this chapter is the one constructed by Brookes and Baker (2021) for their analysis of how obesity is discursively represented in UK newspapers. In particular, we describe the processes and critical considerations that were involved in designing and constructing this relatively large (approximately 36 million words), purpose-built corpus of health-related media language.

3.2.1 Selecting Texts: Representativeness and Balance

When designing a corpus, one of our first considerations is what kinds of texts and language use we want it to represent. This decision, like all considerations underpinning corpus design, will be driven first and foremost by our research questions and/or the broader purpose of our corpus. However, in most cases, we as researchers and corpus builders do not have access to – or even complete knowledge of – the full extent of the texts that could be deemed relevant for our research purposes, even if these purposes are very clearly defined.[1] For this reason, most corpora constitute a mere sample of all the possible texts 'out

[1] Exceptions to this are corpora that are designed to represent so-called closed text types, where all the possible texts are known to us (e.g., if we were studying the published works of a deceased author).

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

there' that could have been included. We will therefore typically have to decide how we are going to sample texts from all the possible language use 'out there' in the world that could suit our purposes and help us answer our research questions. Designing a corpus essentially involves developing a sampling frame to help us decide which texts will be included in the corpus and whether we will include these texts in their entirety or sample material from them (Biber, 2004: 174).

At this point, considerations around representativeness become relevant. Biber (1993: 244) defines representativeness as 'the extent to which a sample includes the full range of variability in a population'. Here the term 'population' refers not necessarily to a group of people but more conceptually to the 'notional space within which language is being sampled' (McEnery and Hardie, 2012: 8). Broadly speaking, the more representative of the target population our sample (or corpus) is, the greater confidence we can have that our findings can be generalised to the population under study.

In assembling a corpus of news reporting about obesity, Brookes and Baker (2021) had to decide what they wanted their data to represent and then design their corpus accordingly. They wanted to assemble British articles that covered the topic of obesity, so they searched for articles using the LexisNexis online news archive, which stores digitised versions of online and print editions of newspaper articles from a range of countries, including the UK. Brookes and Baker (2021) searched for all articles containing one or more mentions of the word *obese* and/or *obesity*. This is a relatively liberal search criterion, compared to corpus studies of news texts more broadly (which often stipulate that search terms need to occur multiple times and/or within the headline or lead paragraph; e.g., Brookes, 2023). Brookes and Baker's more liberal search criteria had the advantage that it yielded a larger number of results. A potential disadvantage of this approach is that it yields a considerable number of articles that are not 'about' obesity but which, rather, mention it in passing. (For a discussion of 'aboutness', see Scott and Tribble (2006); for ways of measuring aboutness in the process of corpus design, see Scott (2017).) Yet, viewed another way, their decision to include such cases brought the analytical advantage that it allowed the researchers to consider instances in which obesity was, in being mentioned in passing, accordingly 'topicalised' within the context of reporting around other news topics and events. It could also be argued that every mention of a topic, even briefly, will contribute towards the incremental effect of discourse around that topic.

As noted, Brookes and Baker (2021) decided to focus on contemporary UK coverage on obesity, and they defined 'contemporary' as the last 10 years prior to the start of their research. For this reason, they searched for articles published between 2008 and 2017 (inclusive). Had they wanted to adopt a more historical focus, they would have had to sample texts from further into the past. This

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

would also have influenced how they went about sourcing texts for the corpus; while LexisNexis provides reliable coverage of recent years, this coverage becomes patchier as we go further back in time – meaning that this repository is not ideally suited for researching historical news texts.

Even once the researchers had decided that they wanted to focus on UK coverage, and within that to adopt a contemporary focus, the British press still holds a lot of variation that the authors needed to consider when building the corpus. Newspapers in the UK can be distinguished according to their coverage (i.e., national, regional), their format or 'style' (i.e., broadsheet, tabloid), their political stance (i.e., left-leaning, right-leaning or centrist) and their frequency of appearance (i.e., daily, weekly).

Brookes and Baker (2021) decided to focus on national as opposed to regional coverage (due to time limits on how much they could analyse). They included both broadsheets and tabloids, and newspapers across the political spectrum. In terms of frequency, they focused on daily newspapers but included as part of this each newspaper's Sunday, online and so-called sister editions. For example, when collecting texts from the *Daily Mail*, they included texts from the *Mail Online* and *Mail on Sunday*. This marked an important way in which Brookes and Baker sought to expand on previous research on the topic, which had focussed, in the main (and in the context of the UK, exclusively), on printed newspapers only. Table 3.1 gives a breakdown of the article and word counts for each newspaper in the corpus, as well as the mean article length for each newspaper.

As this table shows, Brookes and Baker's corpus was not equally balanced in terms of the number of articles and words contributed by each newspaper. One newspaper, the *Mail*, contributed around 30 per cent of the words in the

Table 3.1 *Number of articles, words, and mean article lengths in Brookes and Baker's (2021) corpus*

Newspaper	Articles	Words	Mean article length (in words)
Express	5,193	3,265,741	629
Guardian	5,008	5,238,062	1,046
Independent	4,336	3,303,269	762
Mail	12,805	11,890,340	929
Mirror	3,398	2,202,323	648
Morning Star	152	63,641	419
Star	1,072	370,818	346
Sun	2,286	1,082,808	474
Telegraph	5,680	4,804,351	846
Times	3,948	3,831,868	971
Total	**43,878**	**36,053,221**	**822**

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

corpus, while the contribution of the *Morning Star* is marginal. Average article length also varied considerably, with the *Guardian*'s average being more than double that of tabloids such as the *Star* and *Sun*. Such imbalances are important to acknowledge, and researchers can make adaptations to their sampling frame to regularise the contributions coming from different sources. (See Baker (2009) for a discussion of the design of the Brown family of reference corpora.) Rather than take, for example, 500-word samples from articles, Brookes and Baker (2021) decided to include all texts in their entirety, on the basis that the functions of the representational discourses they identified could be generated across the full length of the text(s) in question. With respect to the imbalance in the number of articles from each newspaper, they argued that this represented the real-life press landscape and, more specifically, the corresponding imbalance in terms of how much obesity coverage is provided by each of the newspapers they sampled. This imbalance did not pose much of an issue for their analysis, since Brookes and Baker (2021) compared different newspapers and sections of the press using relative rather than raw frequency information. Nevertheless, it is important to acknowledge that in such cases, what can be determined from the corpus is based on the language used in some newspapers that are represented more than others. Accordingly, they modulated any claims based on their analysis of the whole corpus with this imbalance in mind.

Once the relevant texts have been identified, researchers can structure their corpus in such a way as to facilitate queries and comparisons at various levels. Brookes and Baker (2021) downloaded and stored the articles in their corpus in a way that would allow them to discriminate at the level of the text. This allowed them to carry out comparisons of the articles according to their format (i.e., broadsheets versus tabloids), political leaning (i.e., left-leaning versus right-leaning), and date of publication (i.e., looking at change over time). In addition to having analytical utility, this method of storing data also allowed Brookes and Baker to assess the balance and representativeness of their corpus (and, in turn, to modulate their analytical claims), in a way that would not have been possible (or at least practical) had they stored their corpus data in a less-organised way (e.g., as a single file with all articles stored together, indistinguishable in terms of newspaper, date of publication, and so on).

3.2.2 Preparing Texts for Corpus Analysis

Along with attending to theoretical considerations, Brookes and Baker (2021) also had to take into account some practical considerations when designing and building their corpus. In this case, the news texts they downloaded from

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

LexisNexis required processing prior to analysis; otherwise they would contain features which can adversely influence the accuracy of analytical procedures.

One such issue, and one that frequently arises in the compilation of corpora of online resources, relates to the presence of so-called boilerplate text. This refers to language that occurs within a text, but which is likely to constitute 'noise' in the context of an analysis. When downloading news articles from LexisNexis, the text files will include labels which indicate, inter alia, the 'headline', 'byline' and the author's name. As such elements occur in every text in the corpus, these can accumulate quickly and thus become problematic for frequency-based analytical measures. In this case, the researchers could rely on the technical support of Andrew Hardie, also working within the Centre for Corpus Approaches to Social Science (CASS), who repurposed these elements as metadata for storing and searching the corpus on the online corpus query processor CQPweb (Hardie, 2012), while rendering such information in a way that would not interfere with corpus querying procedures. Alternatively, they could have used something like the 'boilerplate removal' function of WordSmith Tools (Scott, 2016), which is also much faster than manually removing boilerplate text by hand. Of course, deciding on what counts as 'boilerplate' material is a subjective judgement which has to be made by the researcher and will depend on the aims of the research.

Another practical issue that Brookes and Baker (2021) had to deal with, and one which can be particularly troublesome during the collection of news texts from archives, was the presence of duplicate material. The LexisNexis database can store multiple versions of a single news text (e.g., the online and print versions of an article). This issue was exacerbated further in the case of online articles which reported on developing news stories. This is because such articles constitute 'live' news texts which are typically updated throughout the day as the story develops or corrections to the original story are made. With each update, a 'new' article is essentially produced, and the original article and all subsequent updates were stored as separate texts in LexisNexis. Like boilerplate text, duplicated material can skew the results of frequency-based analytical techniques while also hindering the representativeness of the corpus in a more general sense.

There are a series of steps through which duplicate texts in a corpus can be identified; first, Brookes and Baker (2021) noticed a large number of duplicated articles by manually reviewing chronologically ordered concordance lines. This prompted the second stage, which was to search for long n-grams (e.g., of seven or more words) automatically, using the corpus analysis tool. N-grams of this length typically occur due to the presence of either quoted material or duplicated articles. Third, the authors searched more systematically for such duplicated material using the 'duplicate text' function within version 7 of WordSmith Tools (Scott, 2016). This function allows users to group the texts

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

within a corpus according to a user-determined threshold of linguistic similarity and then manually check and, if they wish, remove high-similarity results. On the whole, this approach was effective for Brookes and Baker (2021); however, there is no entirely reliable automated way of identifying and removing duplicate texts, and manual checking at each stage is advised. Moreover, like the identification of boilerplate material, deciding on what counts as a duplicate text is up to the corpus compiler and will depend on the aims of the research. For example, to allow for the presence of the repeated use of press copy by the different news providers represented in their corpus (PA Media, Thomson Reuters, United Press International), Brookes and Baker (2021) only searched for and removed duplicated texts *within* newspapers (including online, Sunday and 'sister' editions), rather than looking across newspapers.

This example, from Brookes and Baker, provides some key considerations for building a new corpus of texts that have not been necessarily created for linguistic research. In the next section, we discuss how researchers might adapt texts that have been collected for research purposes, though not originally conceptualised as corpus data.

3.3 Working with Pre-existing Transcript Data from Healthcare Settings

Health communication data can be difficult to collect, and researchers have taken opportunities to use existing datasets for conducting additional analyses of the data collected for a prior study. Two of the reported benefits of this type of secondary analysis include (i) overcoming some of the practical restrictions and resource costs associated with carrying out new data collection and (ii) maximising the value of research that has already been collected (Ruggiano and Perry, 2019). Given that much health communication research is necessarily interdisciplinary, there is also a high probability that datasets are investigated for multiple kinds of analysis. However, data collection will likely be conducted according to specific research practices, reflecting the conventions of the field in which the respective contributors work, and this may deviate from what other collaborators on a project are used to. Even within linguistics, there are different ways of working which shape the fundamental elements of research design, such as how data is collected and how it is documented.

The second case study in this chapter focuses on an international collaboration investigating the interactions that take place in Australian Emergency Departments and, specifically, the work involved in operationalising a collection of transcripts as a corpus. In this example, the researchers were given access to a dataset that was intended for manual qualitative analysis and, subsequently, were tasked with reformatting the data for use with modern

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

corpus analysis software. As a result, we offer some guidance on preparing transcripts that can subsequently be used as corpus files, based on what was involved in this reformatting process.

3.3.1 Communication in Emergency Departments

The data analysed in this case comes from Emergency Department (ED) interactions in Australian hospitals and was collected by a team led by Professor Diana Slade from the Institute for Communication in Health Care (ICH), based at Australia National University. The data was collected for the purposes of a study combining 'discourse analysis of authentic interactions between clinicians and patients; and qualitative ethnographic analysis of the social, organisational, and interdisciplinary clinician practices of each department' (Slade et al., 2015: 11).

EDs are a high-stakes interactional health context in which effective communication can result in life-changing outcomes for patients. There is a clear contribution to be made by linguists in helping to ensure that such communication is effective and supports both the health practitioners in carrying out their care duties and the patients in clearly communicating their needs so that they can be met. Accessing authentic ED interactions is difficult, given the urgency of the encounters and the highly sensitive nature of the personal experiences being discussed. Slade and colleagues (2015: 1–2) summarise what they collected as

> communication between patients and clinicians (doctors, nurses and allied health professionals) in five representative emergency departments in New South Wales and the Australian Capital Territory. The study involved 1093 h of observations, 150 interviews with clinicians and patients, and the audio recording of patient-clinician interactions over the course of 82 patients' emergency department trajectories from triage to disposition.

The dataset therefore represents 'one of the most comprehensive studies internationally on patient-clinician communication in hospitals' (Slade et al., 2015: 2). The original research team produced invaluable analyses of the data from a discourse-analytic approach, which they reported in their book *Communicating in Hospital Emergency Departments* (Slade et al., 2015). What the ICH team collected certainly constitutes an amount of data that is appropriate to benefit from the procedures of corpus analysis and so seemed fitting for secondary analysis using such tools.

3.3.2 Data Transfer

The process of sharing data can introduce challenges that affect what is available to researchers conducting secondary analysis. For instance, there

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

may be ethical concerns if the original consent provided by participants did not account for extending access to the data to those beyond the original research team. Anonymisation may be required before the data is shared, which means that some details may be missing at subsequent stages of analysis.

The ICH team conducted observations and interviews with clinicians and patients, though researchers at CASS have only worked (so far) with the observations that were documented as 'patient journeys' – that is, the full range of interactions that patients experienced from the moment that they entered the ED until their departure or admission into hospital. This large-scale project generated numerous files (audio recordings, researcher notes, information sheets, transcripts, etc.), and the ICH team was able to share their documents with the CASS team in an anonymised form (which precluded the sharing of audio files). However, while the ICH team refers to 82 patient journeys, following the transfer of all associated documents and files, researchers at the CASS Centre were able to identify only 72 patient journey transcripts. This demonstrates one of the potential pitfalls of managing and transferring large datasets.

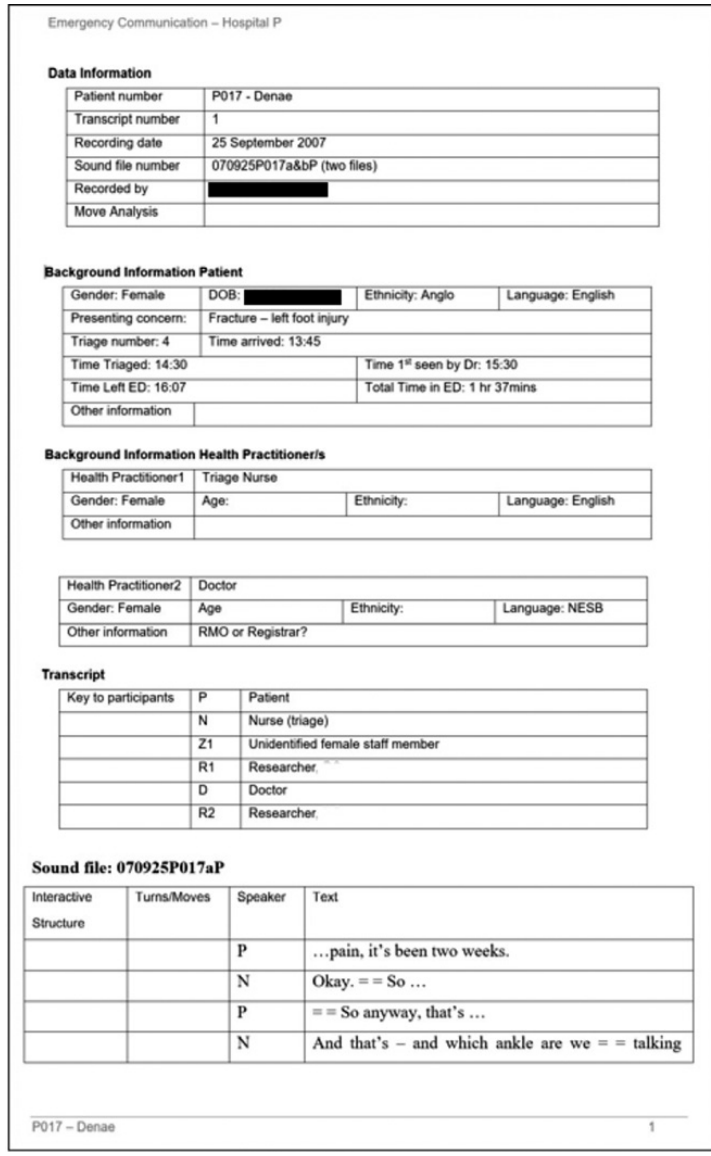
3.3.3 Data Conversion

The ICH team provided word-processed transcripts of the audio-recorded content captured through their observations of the ED encounters. The transcripts were accompanied by metadata about each participant (i.e., role, gender, age, language background, and nationality), and each transcript included a header with information about the context of the 'patient journey' (presenting illness, diagnosis, duration of visit, triage level, number of health professionals seen, and researcher notes), as shown in Figure 3.1.

The process of creating a corpus from this collection of documents involved reformatting the spoken content as plain text and the additional transcriber notes and other non-speech material as annotation using eXtensible Markup Language (XML), which can be computed by corpus analysis tools such as CQPweb (Hardie, 2012) to carry out tokenisation, lemmatisation, part-of-speech (POS) tagging, and semantic tagging.

While it is a common practice in studies of spoken communication to produce a written version, providing researchers with a tangible record that they can analyse and publish in conventional forms of dissemination (which are primarily typed), transcription is inherently a process of recontextualisation and involves decision-making regarding what is documented and how. Not all aspects will be of relevance to the research, resulting in a variety of transcription systems (Richardson et al., 2023: 5). Alongside spoken content, Sperberg-McQueen and Burnard (1994: 250) remind us that 'the production and comprehension of speech are intimately bound up with the situation in which speech occurs',

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003



Emergency Communication – Hospital P

Data Information

Patient number	P017 - Denae
Transcript number	1
Recording date	25 September 2007
Sound file number	070925P017a&bP (two files)
Recorded by	[redacted]
Move Analysis	

Background Information Patient

Gender: Female	DOB: [redacted]	Ethnicity: Anglo	Language: English
Presenting concern:	Fracture – left foot injury		
Triage number: 4	Time arrived: 13:45		
Time Triaged: 14:30		Time 1st seen by Dr: 15:30	
Time Left ED: 16:07		Total Time in ED: 1 hr 37mins	
Other information			

Background Information Health Practitioner/s

Health Practitioner1	Triage Nurse		
Gender: Female	Age:	Ethnicity:	Language: English
Other information			

Health Practitioner2	Doctor		
Gender: Female	Age	Ethnicity:	Language: NESB
Other information	RMO or Registrar?		

Transcript

Key to participants	P	Patient
	N	Nurse (triage)
	Z1	Unidentified female staff member
	R1	Researcher.
	D	Doctor
	R2	Researcher.

Sound file: 070925P017aP

Interactive Structure	Turns/Moves	Speaker	Text
		P	…pain, it's been two weeks.
		N	Okay. = = So …
		P	= = So anyway, that's …
		N	And that's – and which ankle are we = = talking

P017 – Denae 1

Figure 3.1 The first page of a transcript document.

prompting us to incorporate contextual features when we document speech. However, they also assert that 'determining which are relevant is not always simple' (1994: 250). Furthermore, if the researcher plans to investigate the

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

frequency and distribution of such features using corpus tools, they need to be encoded in a format that lends itself to corpus query.

Slade and colleagues (2015: xi) produced orthographic transcripts, favouring standard English spelling but using non-standard forms to capture 'idiosyncratic or dialectal pronunciations (e.g., gonna)'. They refer to fillers and hesitation markers, which were 'transcribed as they are spoken, using the standard English variants (e.g., ah, uh huh, hmm and mmm)' (2015: xi). Finally, they explain that punctuation marks were broadly used according to their meaning 'in standard written English', though also used to indicate additional special meaning, such as putting words in parentheses when the content was unclear and according to 'the transcriber's best analysis' (Slade et al., 2015: xi). Slade and colleagues' (2015) notation system is consistent with wider transcription practices in the field (see Sperberg-McQueen and Burnard, 1994; Fraser, 2022) and demonstrates the influence of the system outlined by Jefferson (2004), which is widely used across the social sciences and strongly associated with conversation analysis. As such, the transcription system aligns with the analytical procedures that the ICH team was to carry out.

We can, however, begin to see potential issues in the transcript notation demonstrated in Jefferson (2004) for corpus tools that rely on computing regular forms. For instance, the Jeffersonian system suggests using parenthesized 'h' to indicate plosiveness (i.e., s(h)orr(h)y). This vocal quality can result from high emotional states, such as crying or laughter, as well as difficulties producing speech because of physical discomfort – all of which would be likely and pertinent in a study of patients in hospital EDs. However, this variant form would not be included in a simple corpus query for 'sorry', which may be an unintended consequence. As an alternative that would also facilitate the automatic conversion to a 'corpus-ready' format, a transcriber could record this as 'sorry [#heavy breathing]'.

Collins and Hardie (2022) summarised the main tasks required in the conversion of the transcript documents to corpus files, and developed recommendations for a transcript notation system that would serve to optimise this process in the future. Collins and Hardie's (2022: 132) recommendations are largely designed 'to generate minimal ambiguity when qualitative-research transcription data is mapped to XML or other structured format and operationalised as a searchable corpus'. An example from the ED data of the kind of ambiguity that such recommendations were created to address comes with the transcriber's use of square brackets, which were used for the following purposes:

- Redacting personal information: 'My name is [patient name].'
- Embedded turns: 'I know it coming, the phlegm black. [D Mm mm] No more.'
- Describing contextual features: [background noise]

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

The alternative suggested by Collins and Hardie (2022: 131) is to indicate different uses of square brackets using a flag character:

- Vocalisation: [@laughs]
- Transcriber comment: [#D fills in form]
- Embedded utterance: [=P yeah]
- Redaction: [name]

This minor adjustment to existing practice will not impede manual analysis but does make it entirely mechanistic to automatically convert the above to XML or another structured format, as follows:

- [@laughs] becomes <voc desc="laughs"/>
- [#D fills in form] becomes <comment content="D fills in form"/>
- [=P yeah] becomes <u who="P" trans="overlap">yeah</u>
- [name] becomes <anon type="person"/>

The designation and function of such flag characters can be adapted to the different levels of specification required for the study.

An extract of one of the transcripts is shown in Table 3.2, and the CASS team was able to use this tabulated format to automatically extract speaker information and spoken content, applying additional scripts to retain the non-speech material that also appeared in the rows in the column (sometimes marked with different formatting; e.g., bold font).

We can see in Table 3.2 how the rows of the table align the spoken material (and sometimes other contextual information) with a speaker ID, and when combined with the case file (i.e., the specific patient journey), this creates a unique speaker identifier (P27_D2), which was recorded in XML as follows:

> <u who=P27D2>And they said you had a little bit of (.) blockage there but nothing too exciting?</u>

This example shows the introduction of utterance boundaries, marked at the beginning (with additional speaker information) and the end, with each cell in

Table 3.2 *Extract of a transcript in the Emergency Department corpus*

Speaker	Text
D2	And you had an angiogram eight years ago was it?
P	Mm, something like that.
D2	And they said you had a little bit of . . . blockage there but nothing too exciting?
P	Well they – because the blockage was close to a branch . . .
D2	Yes.
P	Ah, the surgeon said no, it's too dangerous to operate. (background noise)

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

the 'Text' column, as shown in Table 3.2, approximately mapped onto a speaker turn or utterance. Establishing utterance boundaries and linking these to speakers enabled the research team to subsequently carry out restricted queries and target speakers according to their metadata (e.g., only content provided by doctors and/or female speakers and/or speakers, aged 30–34). This restricted query functionality was used by Collins and co-authors (2022) to investigate the frequency and type of questioning utterances as they were produced by doctors according to different levels of seniority. Finally, the CASS team was able to refer to the transcript notation (with additional manual checking) to separate contextual information and non-speech material from the content in the 'Text' cells and convert them to a format that would exclude them from token counts and content queries, as follows:

<event desc="background noise"/>

Figure 3.2 demonstrates how the extract appeared, following conversion, in the extended context view of CQPweb. This includes the reformatting of pauses to minimise the potential ambiguity of using ellipses, '(. . .)', by using the alternative, more clearly defined notation '(.)'.

As shown earlier in Figure 3.1, the metadata associated with each patient journey was also documented according to a template. The ICH team, however, was working in highly dynamic environments, striving to minimise the disruption to the natural flow of events caused by their presence. As such, as is the case with most large-scale data-collection studies, some details were not recorded and there is variability in the form of the metadata that was collected. For instance, participant ages were indicated to different levels of specificity ('22', 'mid-30s', 'young') and based on the researcher's approximation. The CASS team thus devised a regularised categorisation scheme for age (i.e., 18–24, 25–29, 30–4,

P27_D2: And you had an angiogram eight years ago was it ?

P27_P: Mm , something like that .

P27_D2: And they said you had a little bit of *(.)* blockage there but nothing too exciting ?

P27_P: Well they - because the blockage was close to a branch

(.) **P27_D2:** Yes .

P27_P: Ah , the surgeon said no , it 's too dangerous to operate .

(background noise)

Figure 3.2 A passage of the transcript as it appears in the extended context view of CQPweb.

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

35–9, etc.) and other metadata that allowed us to retain a large portion of the recorded information while also supporting combinations of restricted corpus queries (e.g., selecting 30–4 and 35–9 would facilitate a search of content produced by speakers in their thirties).

We recommend comprehensively documenting contextual information for current and future research projects. However, researchers may be working with data which was produced and archived in the past and for which such information needs to be recovered. In the next section, we discuss what this can involve for corpus compilation.

3.4 Building a Historical Corpus of Anti-vaccination Literature

Our final example in this chapter involves building a corpus of historical texts, specifically the Victorian Anti-Vaccination Discourse Corpus (VicVaDis), which was briefly introduced in Chapter 1 and is discussed by Hardaker and colleagues (2024). In this section we describe what motivated the creation of this corpus and the process that led to its completion. We discuss the decisions that needed to be made along the way, focusing particularly on those that apply to the building of historical corpora. In contrast with the examples presented in the previous two sections, the team who created the VicVaDis corpus (Hardaker et al., 2024) did not initially know whether enough relevant texts had survived from the Victorian period to create a dataset large enough to benefit from the use of corpus methods. Even when the researchers established that a reasonable number of texts were indeed available and convertible to machine-readable format, it was impossible to be certain about how representative those texts might be of the totality of relevant texts from the period, because many, or perhaps most, have not survived until the present day. These issues are also likely to apply to the creation of historical corpora more generally, including from different periods and on different topics.

3.4.1 Vaccine Hesitancy in History

We begin with the motivation for building the VicVaDis corpus. Vaccinations are well known to be the focus of present-day controversies. On the one hand, they are deemed to be one of the most effective public health tools and have made it possible to eradicate or reduce the impact of a wide range of infectious diseases, including smallpox, polio, tetanus, meningitis, yellow fever, measles, Ebola and, most recently, COVID-19 (WHO, 2022). On the other hand, in 2019 the World Health Organization included among the top 10 threats to global health ‘vaccine hesitancy’, which they defined as ‘the reluctance or refusal to vaccinate despite the availability of vaccines’ (WHO, 2019). Indeed, vaccine hesitancy has affected vaccination programmes in 90 per cent of countries in

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

the world (Lancet, 2019) and was further highlighted as a major public health concern during the COVID-19 pandemic (e.g., Hughes et al., 2021).

The spread of anti-vaccination arguments has also been associated with the internet and the rise of social media (e.g., Nuwarda et al., 2022). However, vaccine hesitancy is not in fact a recent phenomenon and has been a concern for as long as vaccinations have existed (Durbach, 2005). The VicVaDis corpus was built to capture the anti-vaccination movement that arose in the second half of the nineteenth century in England, during the reign of Queen Victoria, when the vaccine against smallpox was made compulsory for babies (Tafuri et al., 2014). Smallpox was highly infectious and had a mortality rate of at least 30 per cent (Stewart and Devlin, 2006). At the end of the eighteenth century, Edward Jenner, an English doctor, built on long-standing folk knowledge to assert that those who had contracted the milder illness cowpox were immune to smallpox. He subsequently introduced the first 'vaccination' (from Latin *vacca*) – that is, the practice of introducing material from the pustules of cowpox sufferers under the skin of a healthy person, causing a reaction that resulted in immunity to smallpox. Jenner is deemed to have saved more lives than any other person in history (Stewart and Devlin, 2006), and following widespread vaccination campaigns, smallpox was declared eradicated by the Thirty-Third World Health Assembly in 1980.

In nineteenth-century England, however, mandates for vaccination and the use of material derived from animals were among the reasons that caused major resistance, manifested, for example, in a demonstration involving around 100,000 people in Leicester in 1885 (Charlton, 1983; Durbach, 2005). For insight into the vaccine-hesitant sentiment at the time, consider the following extracts from 2 of the 133 texts included in the VicVaDis corpus (Hardaker et al., 2024: 169–70):

> 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?*)

> A wider, and deeper, and subtler *Social* Evil than universal Compulsory Vaccination is scarcely conceivable; on the physical side, universal pollution; on the side of manhood, womanhood, and childhood, with their several dignities, it is to the extent of its reach, degradation and extinction. The cradle is born to an immediate medical hell. Politically, *Compulsory Vaccination* is an innermost stab of Liberty which piercing its heart, will find its courage and heaven-born principles and convictions in other directions an easy prey. State medicine can do what it likes with us, if we once let it do this. (1879, LSACV, *The Vaccination Inquirer and Health Review: The Organ of the London Society*)

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

The Victorian anti-vaccination movement has been studied by historians (e.g., Durbach, 2005), but prior to the creation of the VicVaDis corpus, there was no textual dataset that would enable large-scale linguistic studies of the anti-vaccination arguments of the time. The corpus was thus created to make it possible to carry out such studies, in order to complement historical research and to enable comparisons with present-day discussions of vaccinations, which already benefit from the availability of corpora (e.g., Coltman-Patel et al., 2022). As we describe in more detail in the following sections, the VicVaDis corpus contains approximately 3.5 million words, drawn from anti-vaccination texts that would have been widely available to the population at the time.

3.4.2 Creating the Corpus

Hardaker and colleagues (2024) began with a series of exploratory searches of online collections of texts covering the relevant historical period to establish how much textual material was still available on the topic of vaccination. This revealed that a substantial amount of such material still existed, particularly in three specific text archives: (1) the Wellcome Collection Library (https://wellcomecollection.org), (2) Project Gutenberg (www.gutenberg.org), and (3) Online Library of Liberty (https://oll.libertyfund.org). The team used 'vaccination' as a search term to query all three text archives. Among other things, this led to the retrieval of *A Catalogue of Anti-Vaccination Literature*, which was issued by the *London Society for the Abolition of Compulsory Vaccination* (the LSACV). This catalogue, which contains 205 texts, was used to identify any additional documents beyond those retrieved from the initial three archives. This led to more targeted searches in the UK Medical Heritage Library (UKMHL, www.medicalheritage.org), the British Library Nineteenth-Century Collection (www.bl.uk), JSTOR (www.jstor.org), and the Internet Archive (https://archive.org).[2] Once these documents were retrieved, the criteria for including them in the corpus concerned topic, time, location, genre, and technical quality.

With regard to topic, the team only included texts that took an anti-vaccination stance. This was established by considering titles (e.g., *The Evils of Vaccination: With a Protest against Its Legal Enforcement*), consulting sources (e.g., the

[2] These resources linked to documents in the following further archives: the US National Library of Medicine, Bristol Selected Pamphlets, the London School of Economics Selected Pamphlets, the Francis A. Countway Library of Medicine, the Harold B. Lee Library, the London School of Hygiene and Tropical Medicine Library and Archives Service, the Royal College of Physicians in Edinburgh, the Royal College of Surgeons of England, and the online collections of Harvard University, Oxford University, Saint Mary's College of California, University of California, University of Glasgow, University of Leeds, and the Cushing/Whitney Medical Library at Yale University. In addition, some documents were uploaded by private individuals to the Internet Archive.

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

London Society for the Abolition of Compulsory Vaccination) and, where necessary, becoming familiar with the texts themselves.

With regard to time period, the team considered the dates of the seven vaccination acts that were passed by Parliament in England with regard to the smallpox vaccine. The first such act dated from 1840, but it was the 1853 act that made mandatory the vaccination of babies by the age of three months, with the introduction of fines and imprisonment for parents who did not comply. Four subsequent acts introduced various changes in legislation but did not substantially change the mandatory nature of vaccination. This changed with the 1907 Vaccination Act (shortly after the end of Queen Victoria's reign), which made it relatively easy for parents to opt out of vaccination without incurring penalties. This effectively marked the end of compulsory smallpox vaccination. Against this background, it was decided that the VicVaDis corpus would include documents published between 1854 and 1906.

With regard to place of publication, the researchers were interested in documents from England and Wales, where the vaccination acts applied. However, no documents published in Wales were retrieved from the various archives. Therefore, the VicVaDis corpus includes only texts published in England.

With regard to genre, the searches posed by Hardaker and colleagues identified a wide variety of types of texts that could be considered for inclusion:

- Pamphlets and popular journals produced by anti-vaccination campaigners
- Local newspaper reports and letters to local newspapers
- Papers presented to Parliament by John Simon in 1857 (e.g., cited by Williamson, 2007; Durbach, 2005)
- Letters from prominent doctors to various public bodies and publications
- Medical journals: *The Lancet* and the *BMJ* (analysed by Kondrlik, 2020)
- A summary of positions on vaccination in the *Edinburgh Review* (1810), a leading political and literary magazine (analysed by Williamson, 2007)
- Materials gathered in the *Monthly Review*, which predated the *Edinburgh Review*, aimed at non-specialist but educated readers (analysed by Arnold & Arnold, 2022)
- Contemporary medical histories and textbooks (Hardaker et al., 2024: 165)

While this variety of texts was available for inclusion, the researchers' primary interest was in non-fictional texts that were aimed for and accessible to the general public, namely pamphlets, newsletters, non-academic tracts and periodicals, and letters to newspapers. The corpus thus excludes fiction and drama, and legal, scientific, and medical publications. The reasons for this decision were as follows: Although literacy rates in England and Wales increased rapidly during the nineteenth century, specialist texts aimed for expert educated audiences had limited circulation, due to a combination of cost and venue of publication. They are also the texts that tend to be relied upon

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

in existing historical studies. In contrast, more 'popular' texts have not been extensively studied but were in fact widely available to people at the time. Pamphlets, for example, were produced quickly and distributed cheaply or free of charge (Humphries, 2011). As such, the material included in the corpus can be seen as the closest equivalent at the time of today's posts on social media.

With regard to technical quality, inclusion was based on the success rates of employing optical character recognition (OCR) software to convert the documents retrieved – mostly as PDFs – from the archives into a plain text format that could be processed by (corpus) software tools. Having tested the quality of different OCR tools, the conversion process was carried out via Adobe Acrobat, and a tailor-made Perl script was employed to identify words in the converted documents that included non-alphabetic characters, such as numbers and punctuation. A quality threshold of 70 per cent accuracy was then adopted as a criterion for inclusion in the corpus. In other words, documents with scores of less than 70 per cent were excluded. This was regarded as an adequate reflection of the error rate in the process of conversion.

Finally, in contrast with the creation of many other types of corpora, the VicVaDis corpus did not pose any issues of ethics and copyright. All relevant documents were in the public domain, and any copyright restrictions had long expired.

3.4.3 The VicVaDis Corpus

The VicVaDis corpus consists of 133 texts and 3,488,959 tokens calculated using the corpus analysis toolkit AntConc (Anthony, 2023). Texts vary considerably in length. The shortest is a 195-word letter to the *East London Observer* from 1985. The longest is *The Anti-Vaccinator and Public Health Journal 1872–3*, a 362,864-word anthology edited by John Pickering which includes letters, notes, public addresses, and reports. The texts involve 66 unique declared authorship designations. These include not just named individuals but also, for example, anonyms and pseudonyms (e.g., *A Sufferer*) and organisations (e.g., the LSACV).

The chronological dispersion of documents in VicVaDis is shown in Figure 3.3 (Hardaker et al., 2024: 7).

The mean number of texts per year is 2.5, but as Figure 3.3 shows, the number of texts per year varies, with no texts in some years and peaks in others (e.g., 11 documents in 1889). However, we should not necessarily consider the peaks and troughs as a reflection of the number of anti-vaccination documents published in different years. Rather, these figures are largely dependent on what was retrievable from the various archives, and on the error rate that resulted from OCR conversion.

Overall, for historical corpora such as the VicVaDis corpus, it is particularly difficult to reach the standards of representativeness and balance

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

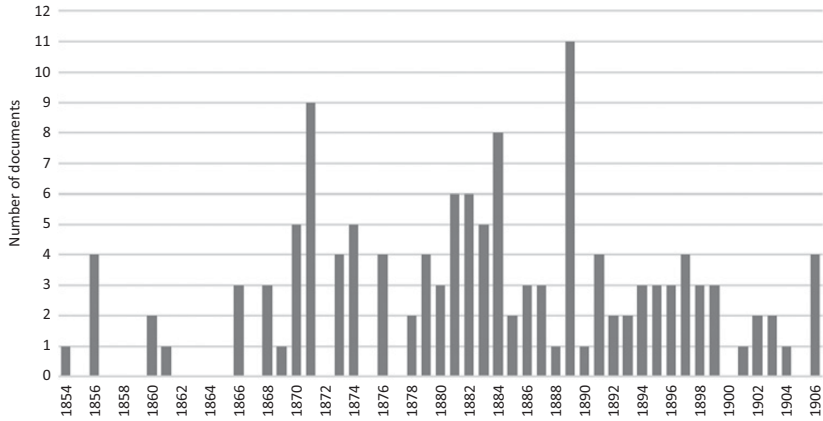


Figure 3.3 Chronological dispersion of texts in the VicVaDis corpus.

that are achievable in the creation of present-day corpora, such as the news corpus on obesity described earlier in this chapter. Hardaker and colleagues (2024) explained this as follows:

> As in any corpus, there is a tension between the human-imposed notion of *balance* across, for instance, authors, texts, dates, and so forth, and the more organic principle of *representativeness*.. . . Even in contemporary corpora, objective representativeness is generally an ideal, and in historical corpora such a goal is less achievable still, given the fundamental lack of ground truth. In our case, we have no way of identifying all of the anti-vaccination literature in circulation throughout the 53 years that our texts cover, and some – or possibly, most – of it will be lost without trace. We therefore have no rigorous benchmarks against which to measure our corpus and so cannot know whether particular years, authors, or texts are over- or under-represented. (Hardaker et al., 2024: 167)

It may in fact be possible to add further texts to the corpus if more documents are retrieved in the future and/or if improvements in the success rates of OCR conversion bring more documents above the 70 per cent technical quality threshold. Nonetheless, as shown in Hardaker and colleagues (2024), even with its inevitable limitations, the VicVaDis corpus constitutes an important resource for the study of anti-vaccination arguments in the Victorian period, and for diachronic or cross-cultural comparisons of vaccination-related discourses. The corpus is freely available through the UK Data Service (https://reshare.ukdataservice.ac.uk/856736/).

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

3.5 Conclusion

Researchers are tasked with making a number of decisions when it comes to collecting data, and corpus linguists typically have to reconcile the theoretical principles of good corpus design with the practical challenges of data availability. The resources through which we access our data can be very influential in shaping the corpus that we eventually compile, both in terms of what is included in a data archive, for example, and the quality of the content and associated metadata. What we have shown through the case studies discussed in this chapter is that however we come to collect our data, there are certain steps that we can take to optimise the utility of our corpus for systematic and rigorous analysis. For instance, we can reduce the 'noise' of our investigations of health discourses by removing 'boilerplate' text, and we can make sure that our interpretations are informed by contextual factors by recording salient metadata. Taking the time to pre-process the data, for example, in applying annotation for paralinguistic features or regularising spelling, can also be beneficial for enabling researchers to query these features. Ultimately, familiarising oneself with the characteristics of the data and the capabilities of the corpus software tools can ensure that researchers are able to get the most out of the data.

References

Anthony, L. (2023). *AntConc* (Version 4.2.4) [Computer Software]. Waseda University. Available from www.laurenceanthony.net/software.

Arnold, W. and Arnold, C. (2022). Medicine in the *Monthly Review*: Revealing Public Medical Discourse through Topic Modelling. *Digital Scholarship in the Humanities*, *37*(3), 611–29. https://doi.org/10.1093/llc/fqab034.

Baker, P. (2009). The BE06 Corpus of British English and Recent Language Change. *International Journal of Corpus Linguistics*, *14*(3), 312–37. https://doi.org/10.1075/ijcl.14.3.02bak.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–57. https://doi.org/10.1093/llc/8.4.243.

(2004). Representativeness in Corpus Design. In G. Sampson and D. McCarthy (eds.), *Corpus Linguistics: Readings in a Widening Perspective* (pp. 174–97). Continuum.

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.

Charlton C. (1983). The Fight against Vaccination: The Leicester Demonstration of 1885. *Local Population Studies*, *30*, 60–6. Available at www.localpopulationstudies.org.uk/PDF/LPS30/LPS30_1983_60–66.pdf.

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

Collins, L. C., Gablasova, D. and Pill, J. (2022). 'Doing Questioning' in the Emergency Department (ED). *Health Communication*, *38*(12), 2721–9. https://doi.org/10.1080/10410236.2022.2111630.

Collins, L. C. and Hardie, A. (2022). Making Use of Transcription Data from Qualitative Research within a Corpus-Linguistic Paradigm: Issues, Experiences and Recommendations. *Corpora*, *17*(1), 123–35. https://doi.org/10.3366/cor.2022.0237.

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.

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

Fraser, H. (2022). A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. *Frontiers in Communication*, *7*, 898410. https://doi.org/10.3389/fcomm.2022.898410.

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.

(2014). Modest XML for Corpora: Not a Standard, but a Suggestion. *ICAME Journal*, *38*(1), 73–103. https://doi.org/10.2478/icame-2014-0004.

Hughes, B., Miller-Idriss, C., Piltch-Loeb, R., Goldberg, B., White, K., Criezis, M. and Savoia, E. (2021). Development of a Codebook of Online Anti-Vaccination Rhetoric to Manage COVID-19 Vaccine Misinformation. *International Journal of Environmental Research and Public Health*, *18*(14), 7556. www.mdpi.com/1660-4601/18/14/7556.

Humphries, B. (2011). Nineteenth Century Pamphlets Online. *The Ephemerist*, 153. The Ephemera Society. Available at http://eprints.lse.ac.uk/40202/.

Jefferson, G. (2004). Glossary of Transcript Symbols with an Introduction. In G. Lerner (ed.), *Conversation Analysis: Studies from the First Generation* (pp. 13–31). John Benjamins. https://doi.org/10.1075/pbns.125.02jef.

Kondrlik, K. E. (2020). Conscientious Objection to Vaccination and the Failure to Solidify Professional Identity in Late Victorian Socio-Medical Journals. *Victorian Periodicals Review*, *53*(3), 338–71. https://doi.org/10.1353/vpr.2020.0032.

The Lancet Child & Adolescent Health. (2019). Vaccine Hesitancy: A Generation at Risk. *The Lancet Child & Adolescent Health*, *3*(5), 281. https://doi.org/10.1016/S2352-4642(19)30092-6.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003

Nuwarda, R. F., 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.

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.

Reppen, R. (2022). Building a Corpus: What Are Key Considerations? In A. O'Keeffe and M. J. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*, 2nd ed. (pp. 13–20). Routledge. https://doi.org/10.4324/9780367076399-2.

Richardson, E., Hamann, M., Tompkinson, J., Haworth, K. and Deamer, F. (2023). Understanding the Role of Transcription in Evidential Consistency of Police Interview Records in England and Wales. *Language in Society.* Online first. 1–32. https://doi.org/10.1017/S004740452300060X.

Ruggiano, N. and Perry, T. E. (2019). Conducting Secondary Analysis of Qualitative Data: Should We, Can We, and How? *Qualitative Social Work*, *18*(1), 81–97. https://doi.org/10.1177/1473325017700701.

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

(2017). *News Downloads and Aboutness*. Plenary speech at the International Corpus Linguistics conference. Birmingham: University of Birmingham. 26 July 2017. www.youtube.com/watch?v=3FVa0KwtvLc.

Scott, M. and Tribble, C. (2006). *Key Words and Corpus Analysis in Language Education*. John Benjamins.

Slade, D., Manidis, M., McGregor, J., Scheeres, H., Chandler, E., Stein-Parbury, J., Dunston, R., Herke, M. and Matthiessen, C. M. I. M. (2015). *Communicating in Hospital Emergency Departments*. Springer.

Sperberg-McQueen, C. and Burnard, L. (1994). *Guidelines for Electronic Text Encoding and Interchange (TEI P3). Text Encoding Initiative.* https://tei-c.org/Vault/GL/p4beta.pdf.

Stewart, A. J. and Devlin, P. M. (2006). The History of the Smallpox Vaccine. *Journal of Infection*, *52*(5), 329–34. https://doi.org/10.1016/j.jinf.2005.07.021.

Tafuri, S., Gallone, M. S., Cappelli, M. G., Martinelli, D., Prato, R. and Germinario, C. (2014). Addressing the Anti-Vaccination Movement and the Role of HCWs. *Vaccine*, *32*(38), 4860–5. https://doi.org/10.1016/j.vaccine.2013.11.006.

Williamson, S. (2007). *The Vaccination Controversy: The Rise, Reign and Fall of Compulsory Vaccination for Smallpox*. Liverpool University Press.

World Health Organization (WHO) (2019). Top Ten Threats to Global Health in 2019. www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019.

(2022). Vaccines and Immunization. www.who.int/health-topics/vaccines-and-immunization.

Downloaded from https://www.cambridge.org/core. IP address: 18.97.14.88, on 09 Sep 2026 at 09:27:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.003