2 Research Questions

2.1 Introduction

Research projects generally involve the formulation of questions that researchers aim to answer via the most appropriate combination of data and methods. In many disciplines, research questions are a requirement in students' dissertation/thesis proposals and in researchers' applications for funding. In practice, research projects do not always begin with specific questions, and even when they do, those questions often evolve over time. However, the development of research questions tends to happen early in the research process, which is why we consider this topic here in this second chapter.

Broadly speaking, research questions need to be relevant and viable in the context where they are intended to be answered. However, exactly what counts as an appropriate research question varies from discipline to discipline and context to context. With regard to projects in linguistics, Sunderland (2010) and Wray and Bloomer (2021) provide useful reflections and guidance. In this chapter we draw from our experience to focus more specifically on the different ways research questions can be developed in corpus-based studies of health communication. We have selected three case studies that contrast with each other in terms of when and how research questions were formulated and how much control we as linguists had in that process.

The first case study involves the analysis of a corpus of UK news articles about obesity. Here the researchers had a considerable degree of freedom in how to approach the study and, specifically, in terms of when and how to formulate research questions. We show how the researchers began with an initial exploratory approach to the data by means of a keyness analysis. They then went on to develop goals and priorities for the research in an organic, cyclical manner that also involved literature reviews and interactions with stakeholders. By these methods, they arrived at a set of specific research aims which, for some purposes, were expressed as a series of research questions.

The second case study involves the exploitation of an existing corpus of English to investigate potential weaknesses in the language used in a diagnostic questionnaire for pain. Here the linguists involved in the research were

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

approached by a pain clinician who wanted to know why her patients seemed to have difficulties with some specific aspects of the pain questionnaire. The researchers then turned the clinician's broad question into a series of specific research questions that could be answered by means of corpus linguistic methods. We show how this made it possible to identify some aspects of the language used in the questionnaire that explained the difficulties observed by the clinician and that also had wider relevance for any health professional using the questionnaire.

The third case study involves the analysis of a corpus of patients' online feedback on the services of the National Health Service (NHS) in England. Here the linguists involved in the research were approached by NHS England and provided with a set of 12 pre-formulated questions that the researchers had to answer in a very tight timeframe. We show how the researchers answered these questions by creatively and eclectically employing the corpus linguistic tools most appropriate for each question. We also point out how some of the original questions had to be adapted in interaction with the external partners and how some additional questions were formulated in response to initial findings. Eventually, all 12 preset questions were answered to the satisfaction of NHS England and within the required timeframe.

Throughout this chapter, we discuss both the challenges and the opportunities associated with each of these different approaches to the development of research questions.

2.2 Developing Exploratory Questions

Perhaps one of the most common ways of approaching a corpus-assisted discourse analysis project is to develop and refine questions as a result of exploratory analyses in a bottom-up manner, an approach which draws on grounded theory (Glaser and Strauss, 1967). Rather than beginning with a specific set of questions, the analyst approaches the corpus in a reasonably naïve and open way, simply asking 'What is interesting, (unexpectedly) frequent, or unusual in this corpus?' and letting the initial answers to those questions lead to further questions. To illustrate this approach, we describe a study (Brookes and Baker, 2021) which involved the analysis of a 36-million-word corpus of newspaper articles about obesity published between 2008 and 2017, with articles drawn from 11 national UK newspapers. Brookes and Baker chose to carry out this analysis by engaging with existing (non-corpus) research which had highlighted problematic aspects of news reporting around obesity; they also drew from their own hypothesis, based on analysing a smaller sample of data, that a corpus approach would be fruitful. They did not begin with any specific lists of words or other linguistic phenomena which they wanted to examine, although they decided to devise a set of possible ways to approach the analysis.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

One approach was to apply some form of comparative analysis. As the corpus contained articles from 11 newspapers across 10 years, 2 obvious forms of analysis were selected. The first was a comparison across newspapers. Based on a previous comparative approach to a corpus of articles about Islam (Baker et al., 2013), Brookes and Baker (2021) decided that making distinctions between 11 newspapers would not be appropriate, especially because some newspapers contributed much smaller amounts of corpus data than others. Instead, they carried out a four-way comparison, based on the tabloid versus broadsheet formats and left versus right political perspectives. They grouped the *Express*, *Mail*, *Star*, and *Sun* into right-leaning tabloids, while the *Guardian*, *Independent*, and *i Paper* were considered together as left-leaning broadsheets. Keyword comparisons between these sets of newspapers highlighted the major lexical differences and similarities among them. Additionally, the researchers considered cases where a single newspaper contributed towards the majority of instances of a specific keyword.

As the corpus consisted of a decade of articles, Brookes and Baker also considered change over time, taking each year of data separately and tracking lexical changes over time. This enabled them to identify how the newspapers gradually moved towards emphasising personal responsibility and biological frames around obesity, while de-emphasising the societal frame. They also took a different perspective on change over time (see Chapter 7) by considering the annual news cycle consisting of 12 months. They then compared the articles published in January (across all years) against articles from February and so on. This approach was inspired by Anna Marchi's PhD research, which looked at a single year of the *Guardian*. In her thesis she writes:

> Firstly it should be noted that there is no particularly good reason to choose a calendar year as unit, but it is purely a matter of cultural habit: most societies, in fact, regulate their existence and its interpretation following what Bettini calls 'the power of the calendar' (1995: 21, my translation). A year span was therefore chosen for reasons of convention and the span was limited to one complete and continuous year of the newspaper's life, in order to limit the impact of the diachronic variable. (Marchi, 2014: 15)

The analysis carried out by Brookes and Baker which compared months was able to show how stories about obesity operate in a repetitive annual cycle, with different topics and discourses occurring at various points throughout the year. For example, in January there was a focus on starting a new diet and joining a gym, whereas during the summer months there were articles relaying concerns about being seen in swimwear while on holiday.

Another aspect of the analysis was inspired by the researchers' engagement with non-linguistic literature on obesity, particularly published work around gender, health, and the body (e.g., Bordo, 1993; Gill et al., 2000; Gough, 2010).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

Other aspects were inspired by conversations with experts from other disciplines and stakeholders, where researchers were encouraged to view discourses around obesity through a lens of social class. This led them to carry out analyses which focussed on different kinds of social actor representation in the corpus of articles: men and women, as well as the terms *under-class*, *working class*, *middle class*, and *upper class*.

By engaging with these initial comparative studies, the analysis helped identify other aspects within the corpus which felt ripe for more detailed study. One of these was the presence of shaming and stigmatising language, which had been identified at an early stage in the research as appearing more clearly in the right-leaning tabloids (e.g., through noun labels like *fatty* and *hog*, adjectives like *lardy* or *blobby*, or verbs like *guzzle* and *waddle*), although more subtle uses of stigmatising language had also been identified in broadsheet newspapers (e.g., the nomination *the obese*). Although stigma was not too frequent in the news, it felt like a salient theme and also one of the most problematic aspects of the articles, from a critical perspective. Therefore, this was deemed worthy of a separate analysis, which Chapter 10 of this book describes in detail. Generally, the phenomenon of stigmatising language has been of particular interest to charities and other groups outside academia.

Finally, the researchers decided to focus part of the analysis around four specific words which they had identified as highly frequent or having a high keyness score in the corpus, as well as collocating with one another. These words were *healthy*, *body*, *diet*, and *exercise*. All four words were significant in that they were used in articles which focussed on different ways of reducing obesity, although they were also used in a wide range of ways, indicating that different meanings and discourses were realised through them.

It is notable that throughout the monograph based on the analysis of this corpus (Brookes and Baker, 2021), there is only one explicit mention of 'research questions', and there is no place in the book where all the research questions are listed, as in Baker and colleagues' (2019) book on NHS feedback, described later in this chapter. Instead, at the start of each chapter of the book, the researchers outlined the topic they aimed to *explore* (e.g., forms of this verb occur 35 times across the book, and it is a word used particularly often in the conclusion chapter). However, in giving conference presentations on various aspects of the research, the researchers summarised what they did in a slide entitled 'Research Questions', with a set of questions that were retrospectively fitted to the analyses that were carried out. So, for example, a presentation which focussed on stigmatising and change over time contained an early slide with the following questions:

1. How do different newspapers represent people with obesity?
2. What legitimation strategies are used with negative representations?

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

3. Has stigmatising language decreased over time?
4. In what other ways has discourse around obesity changed over time?

There is considerable freedom in naming the research questions at the end of a project, along with allowing them to develop organically through a combination of reading around the topic, conversations with others, exploratory corpus procedures, or simply following hunches or analytical paths that look potentially interesting. However, as we will see in the rest of this chapter, there are other ways to develop research questions, which may bring with them unforeseen advantages.

2.3 Developing Questions in Interaction with Stakeholders

In this section we turn to a project where interactions with healthcare practitioners led to the formulation of research questions that could be answered by means of corpus linguistic methods. In this case the focus was a language-based questionnaire for the diagnosis of pain: the McGill Pain Questionnaire (MPQ; Melzack, 1975). To contextualise this project, we start by providing some background on pain, the diagnosis of pain, and the role of language within it. We then introduce the MPQ and the issues associated with it that led to the formulation of research questions which were suitable for a corpus linguistic approach, as part of a collaboration between linguists and a pain clinician (Semino et al., 2020).

Pain can have a wide variety of causes. A fundamental distinction can, however, be made between *nociceptive* and *neuropathic* pain. Nociceptive pain is caused by damage to bodily tissues, as in the case of cuts, burns, and fractures (Vadivelu et al., 2011). As such, nociceptive pain is arguably the most 'prototypical' kind of pain. Other things being equal, it is also relatively straightforward to diagnose, as it is possible to identify its cause by observation through the naked eye or medical tests such as X-rays or scans. In contrast, neuropathic pain is not, or not only, the result of bodily damage but is caused by problems in the nervous system that may not be easily observable, including via X-rays or scans (Wilkie et al., 2001). Neuropathic pain also tends to become chronic (i.e., to last more than three months). The most extreme example of neuropathic pain is phantom limb pain, which is experienced in a limb that the person no longer has (e.g., following amputation). But more common types of pain, such as headaches and back pain, can have a neuropathic component and thus be difficult to diagnose and treat. The *way* in which the patient describes their pain– what it feels like (its 'quality') and how bad it is (its 'severity' or 'intensity') – is always important in healthcare settings, but it is particularly crucial in the case of neuropathic pain, especially when it is chronic.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

Expressing pain in language is, however, notoriously difficult (e.g., Scarry, 1987). In English, for example, the set of lexical items that have literal meanings relating to pain is relatively small and non-specific (e.g., *pain*/*painful*, *hurt* as noun and verb, *sore*, and *ache/aching*). Consequently, pain is often expressed figuratively. Neuropathic pain particularly tends to be expressed metaphorically in terms of causes of damage to the body (Semino, 2010). For example, we talk about a 'burning pain' in the stomach when there is no fire or a 'splitting headache' when our head is not split.

Against this background, clinicians have developed language-based tools for the diagnosis of pain, such as the MPQ, which is reproduced in Figure 2.1. The MPQ was developed at McGill University in Canada in the 1970s. As the figure shows, it includes 78 possible English linguistic descriptors of pain divided into 20 groups, each consisting of between 2 and 7 descriptors. The division into groups is central to the goal of the questionnaire, which was to capture both the *quality* and *severity* of the patient's pain (i.e., what the pain feels like and how intense it is). The groups capture different qualities or types of pain and fall into four broader classes, depending on the aspect of pain they reflect: sensory (groups 1–10), affective (groups 11–15), evaluative (group 16), and miscellaneous (groups 17–20). Within each group, the descriptors are listed in order of increasing severity or intensity of pain. For example, group 3 captures the sensory quality of *punctate pressure* and contains five descriptors: *pricking*, *boring*, *drilling*, *stabbing*, and *lancinating*. *Pricking* is the descriptor associated with the lowest severity within the group, and *lancinating* is associated with the highest intensity. Group 3 is also one of several groups of descriptors that consist of metaphorical descriptions of the quality of pain in terms of different causes of damage to the body.

When completing the questionnaire, patients have two options with regard to each group: they may not pick any of the descriptors, if that quality of pain does not apply to them (e.g., if their pain does not feel hot, they do not pick any descriptors from group 7); alternatively, if the relevant quality of pain does apply to them, they pick *one* descriptor (namely, according to the design of the questionnaire, the one that best captures the severity of that kind of pain). In this way, by looking at a patient's selections, the clinician has an overview of both the kind(s) of pain that the patient experiences and their intensity.

A few years prior to the writing of this book, one of the authors (ES), who has an interest in communication about pain, was asked for advice about the MPQ by a pain clinician (Dr Joanna Zakzrewska) who regularly employs the questionnaire in her consultations with patients, alongside other approaches aimed at diagnosing the cause of their pain. The clinician reported that her patients sometimes struggled with some of the descriptors included in the MPQ and/or found it difficult to select a single descriptor from the groups that applied to their experience of pain. Indeed, to avoid multiple selections from each group,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

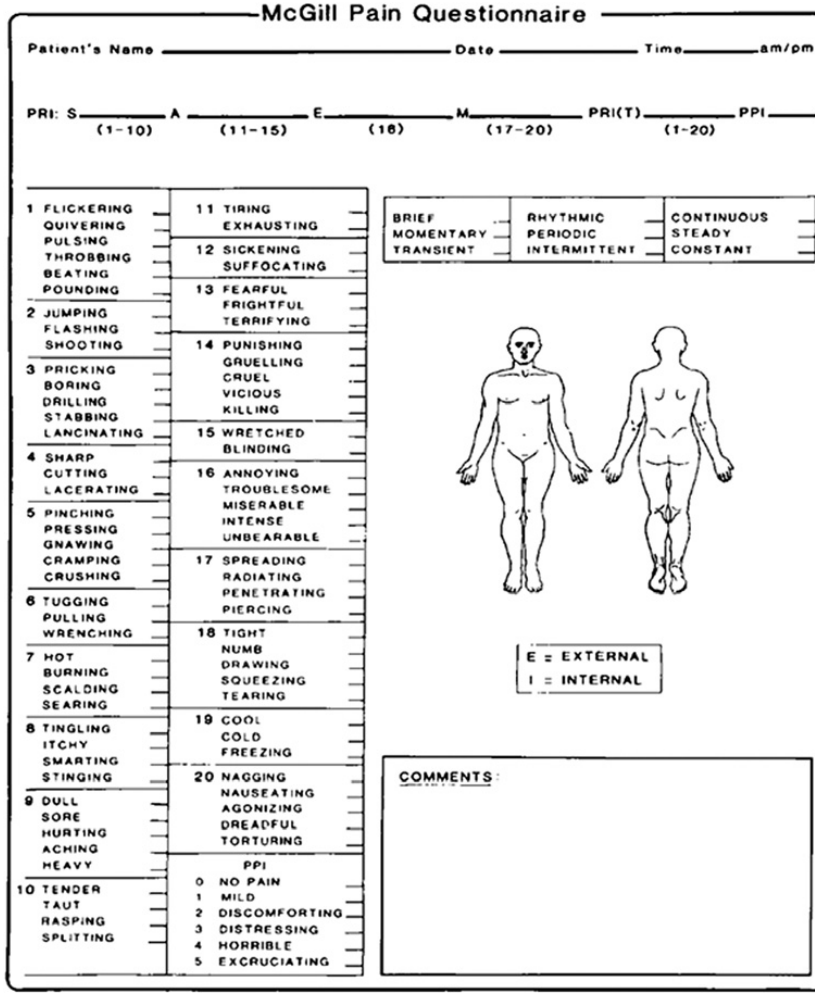


Figure 2.1 The McGill Pain Questionnaire. (MPQ; Melzack, 1983: 44)

this clinician administered the MPQ verbally (i.e., reading out each group to the patient and asking them to select one descriptor only per group). The clinician therefore wondered whether a linguistic perspective on the MPQ might explain the problems encountered by her patients.

As we have explained, the MPQ relies on two dimensions of variation between the 78 descriptors: across groups, there is intended to be variation in terms of pain

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

quality; within each group, there is intended to be variation in terms of pain severity. Even before carrying out any analysis, a linguist might expect that the 78 descriptors are likely to contrast in other ways. However, before attempting to turn the clinician's general question into a set of viable research questions for a corpus linguist, previous studies on the MPQ itself needed to be considered.

The MPQ has been widely used in pain diagnosis, both in its original English language version and in translations into at least 26 additional languages. On one hand, its application to a wide range of conditions has shown that it is a valid, reliable, and sensitive tool. According to a systematic review by Main (2016: 1390), 'there is evidence that the MPQ (1) can discriminate between pain conditions, and also capture variation within conditions, (2) is sensitive to change, and (3) is responsive to treatment and can be used as an outcome measure'. On the other hand, a number of studies have pointed out issues with the MPQ descriptors, including that some are rare words in English (e.g., *rasping*), some may be ambiguous (e.g., *boring*, from the *punctate pressure* group above), and some may not often be used to describe pain (e.g., *taut*; Fernandez and Boyle, 2002). Partly as a result of these issues, two shorter versions of the MPQ have been produced since the launch of the original one: Short-Form-MPQ (SF-MPQ; Melzack, 1987) and Short-Form-MPQ-2 (SF-MPQ-2; Dworkin et al., 2009). The first, SF-MPQ, was developed as a less-time-consuming version of the original MPQ and contains 15 descriptors, each rated on a 4-point intensity scale. The second, SF-MPQ-2, contains 7 additional items intended to be relevant to neuropathic pain (i.e., pain caused by problems in the nervous system) and adopts a 10-point scale for pain intensity.

Against this background, it was then possible to identify two main dimensions of variation among the 78 descriptors that would be possible to investigate by means of corpus methods to address the clinician's concern, and to contribute a new linguistic perspective to existing literature:

- the frequency of each descriptor in English, which can be taken as a proxy measure of each word's familiarity for patients;
- the tendency for each descriptor to be used to describe pain, which can be operationalised in terms of the corpus linguistic notion of collocation.

This led to the formulation of the following research questions:

1. To what extent do the 78 descriptors included in the McGill Pain Questionnaire vary in terms of their frequency in general English?
2. To what extent do the 78 descriptors included in the McGill Pain Questionnaire vary in terms of the strength of collocation with the string *pain*?

Answering these questions required the selection of a suitable corpus of English. The Oxford English Corpus (OEC) was selected for this purpose. The OEC includes 2.5 billion words of twenty-first-century English. It is mainly

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

drawn from material collected from the World Wide Web and contains texts from a wide range of genres and domains (e.g., news and media, law, medicine, science, business, fiction, personal blogs). It also includes texts from international varieties of English from different parts of the world (e.g., UK, US, Australia, India, Singapore). As such, it is an appropriate reference corpus of 'general English'. The OEC is accessible and searchable via the corpus manager and text analysis software Sketch Engine (www.sketchengine.eu; Kilgarriff et al., 2014).

Semino and colleagues (2020) show how answering question 1 above revealed substantial variation in the frequency of the MPQ descriptors in the OEC. The most frequent MPQ descriptor, *hot*, occurs 206,291 times in the OEC (84.857 instances per million words), while the least frequent, *lancinating*, occurs only 15 times (0.006 instances per million words) and only appears in medical research articles about the MPQ itself. More generally, answering question 2 identifies a set of 15 MPQ descriptors that occur less than once per million words in the OEC (e.g., *quivering, smarting*, and *taut*) and thus can be considered relatively rare words. This provided the pain clinician with evidence of what words her patients are more likely to find difficult to understand.

Semino and colleagues (2020) also found considerable variation in the strength of collocation between each descriptor and the word *pain* in the OEC. As part of this, they showed, for example, that *sharp* has 986 co-occurrences with the lemma *pain* as a noun (in a window of 5 words to the left and 5 words to the right of the descriptor), while *rasping* has none. More broadly, 24 descriptors were found to have 10 or fewer instances of *pain* within the relevant collocational window in the whole corpus, including *flashing*, *jumping*, *sickening*, and *tugging*. Collocation is a linguistic phenomenon, but it has been hypothesised to reflect psychological associations between words in our mental lexicons (Hoey, 2005), which can lead to 'priming effects', whereby being exposed to one member of a collocational pair leads to faster recognition of the other member of the pair in experimental settings. In the MPQ, however, any such priming effects could be problematic, as the patient's selections within each group of descriptors are taken to reflect the severity of the patient's pain. This led to the formulation of a third research question:

3. To what extent does variation in the strength of collocation with 'pain' within each of the 20 groups in the McGill Pain Questionnaire correlate with patients' selections for each group?

To answer this question, Semino and colleagues (2020) brought together two sources of data: patients' selections in 800 completed questionnaires at the Eastman Dental Hospital in London and information about the strength of collocation between each descriptor and the noun lemma *pain* in the OEC. Using a standard measure of correlation (the Pearson correlation coefficient),

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

they found that for 7 out of 10 sensory groups in the MPQ, patients' choice of descriptor can be explained largely or entirely in terms of the strength of the collocational link from the word *pain* to that descriptor. For example, in group 2, patients overwhelmingly selected *shooting*, which has a much stronger collocational link with *pain* than the other descriptors and, within the MPQ, is at the top of the severity scale. In group 4, patients overwhelmingly selected *sharp*, which also has a much stronger collocational link with *pain* than the other descriptors but, within the MPQ, is at the bottom of the severity scale.

This finding undermines the reliability of the original version of the MPQ for the measurement of pain severity and goes some way towards explaining why, in the experience of the pain clinician mentioned at the beginning of this section, patients may find it difficult to pick just one descriptor from at least some of the groups. As the two short-form MPQs approach pain severity differently (via a numerical score associated with each descriptor), the answer to research question 3 also suggests that these versions of the questionnaires may be more appropriate for that purpose.

In summary, this section has shown how a language-related problem in healthcare can lead to the formulation of research questions suitable for corpus linguistic methods, and how answering these questions can help address the original problem. The process of developing research questions is fairly typical of corpus-based studies of health communication that develop from interactions between healthcare professionals/researchers and linguists. A possible disadvantage of this approach is that the resulting research questions are not driven by the interests or priorities of linguists, and thus they may not lead to major new insights into language or discourse. On the other hand, as we have shown, answering research questions formulated in interaction with stakeholders can result in findings that have immediate practical relevance. In addition, the process of answering such questions can sometimes require some useful adaptation or development of corpus methods themselves. For example, Semino and colleagues (2020) used a two-pronged approach to collocation in the MPQ study. In Chapter 10 we discuss a study of the representation of hallucinatory voices that required the development of an ad hoc corpus linguistic approach to the analysis of social actors in interview data.

The next section further explores the potential advantages and disadvantages of addressing questions raised by practitioners, by presenting a study where the analytical focus is far more strictly governed from the outset.

2.4 Working with a Set of Pregiven Questions from Non-academic Partners

In this section we outline a study (described in more detail in Baker et al., 2019) where the analysts had limited freedom when it came to decisions about the focus

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

of the project and the subsequent direction of the analysis. When working with external agents, it is often the case that the researcher will be required to address a set of predetermined goals, which may be non-negotiable requirements in exchange for access to a particular corpus. There are potential benefits and pitfalls to this kind of relationship: embarking on a piece of new research with a ready-made dataset and a clear set of preset goals could save time, and there is less need to engage with exploratory forms of research in order to identify areas of interest. However, the questions set by people who have not used corpus linguistics methods before might also be difficult to answer, as we began to show in the previous section. They may not be worded in ways that enable research to be carried out appropriately or effectively; in addition, the kinds of questions being asked might overlook other important aspects of the data.

In 2015, members of the Centre for Corpus Approaches to Social Science (CASS) were contacted by a senior member of the Patients and Information Directorate at NHS England. This section of the NHS was involved with analysing patient feedback which had been posted to a website. At the time, a set of almost 29 million words in comments from patients, along with 11.6 million words in responses from NHS providers, was publicly available, consisting of posts made between March 2013 and September 2015. The researchers at CASS were asked if they would consider carrying out a corpus-assisted discourse analysis of these posts, in order to help NHS England make sense of such a large amount of data, as well as to develop methods of analysis which could be shared with staff at NHS England, so that they could analyse large amounts of feedback in the future. A proviso was that NHS England would provide a set of questions, and the team at CASS would be required to produce detailed reports within 18 months to answer these questions.

The researchers were duly presented with 12 questions which had been compiled during a team meeting at the Patients and Information Directorate (CASS members were not present at this meeting). The team admitted that not all of their members had seen the corpus data in advance, so some of them had struggled a little to devise questions. They conveyed that they were also happy for the members of CASS to consider any additional aspects of interest that emerged as they carried out their analysis. The questions that were set by the NHS team are listed as follows:

1. What are the key drivers for positive and negative feedback?
2. What are the key differences in experience across different providers (e.g., acute providers and General Practitioners)?
3. How consistent are the messages within a provider (or site, department ward, if available)?
4. Are the comments consistent with the quantitative ratings/scores?

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

5. What are the main areas of concern / what matters most to patients (e.g., relational or functional aspects of care)?
6. Who is the focus of the concern raised (e.g., individual staff member – nurse, General Practitioners, general/organisational)?
7. What impact has the experience had on the individual posting a comment?
8. What is the 'quality' of the comments provided (e.g., content, length, clarity, relevance, specificity)?
9. Are there any differences by socio-demographic group?
10. What key words within a text might trigger an alert/urgent review?
11. Can the comments be easily categorised (e.g., positive/negative, important/urgent)?
12. What proportion of comments say something along the lines of 'I've already raised this and you've done nothing about it' (i.e., repeat/ongoing concerns)?

Some of the questions are worded in ways that would suggest a yes/no answer (e.g., questions 4, 9, and 11), something which the researchers on this project would tend to avoid in their own research, instead preferring a wording which allows for a more open answer. For example, question 4 could be rephrased as 'to what extent, and how are the comments consistent with quantitative ratings/scores?' Several questions contained comparative aspects (e.g., questions 1, 2, 4, and 9), which are generally well-suited to corpus-assisted techniques of analysis. Some questions referred to more vague criteria, such as question 3, which referred to 'messages'; question 7, which referred to the impact on the individual; question 10, which referred to words that might trigger an urgent review; and question 12, which referred to repeated or ongoing concerns. A potential disadvantage of corpus-based approaches is that it can be difficult to search and retrieve all cases of a variable linguistic item. For example, there are many ways that someone can indicate that they have raised a concern before. And sometimes simply searching for what appears to be the most obvious phrasings will produce little value. For example, the phrases 'I have raised this concern before' and 'I have already raised this' did not occur at all in the patient feedback corpus. In particular, question 10 raised another question: what kind of problems in the NHS would require an alert or urgent review, and would the criteria for this be qualitative (e.g., something terrible happening), quantitative (e.g., something happening often), or both?

When the researchers at CASS received the list of questions, they found some of them to be more challenging than others, and they reasoned that most of these questions would probably not have been ones that they would have asked of the data, when relying on corpus methods. However, perhaps this could be seen as a positive aspect of working with external partners. The

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

partners did not have a sense of what was easy (or not) for a corpus-assisted discourse analysis, so their questions were based on what they felt was important to know, rather than being restricted by considerations regarding what they thought the tools could tell us. And these 'difficult' questions were interesting in that they required the corpus researchers to work outside their comfort zone, to think creatively about the possible ways that they could be addressed.

The CASS researchers read samples of the corpus in order to identify cases where language was used in relation to the more variable phenomena. For example, question 7 asked about the impact of an experience on a patient, and after some experimentation, it was decided to consider impact in terms of the feelings that the patient described, along with expressions of their intentions. This led to the consideration of phrases like 'I will/will not/won't' and 'I feel'. It was found that the former set of phrases collocated with verbs like *change*, *move*, *leave*, *return*, *recommend*, and *forget*. Examination of concordance lines containing these kinds of collocates helped identify the kinds of cases where patients said they were intending to change their provider and under what circumstances (e.g., poor standard of treatment, long wait times, poor staff interpersonal skills, lack of medication availability) and cases where patients said that they would (or not) recommend the provider to others. Examination of collocates of 'I feel' uncovered examples of people describing how they felt 'let down', 'sorry', 'fortunate', or 'safe', and subsequent concordance analyses were able to provide further detail regarding the reasons for these feelings. There are undoubtedly other ways that patients can talk about their intentions and feelings, but the researchers had identified a set of seed phrases that produced reasonably large enough cases for them to be able to conduct an analysis. The solution, then, was not to identify every phrasing but to find frequently used phrasings that were employed by a range of different people and could be taken as reasonably representative.

What the CASS team found interesting about engaging with the NHS feedback was that a relatively simple research question about demographic differences led to the formulation of a related and more complex set of research questions, regarding what happens when people reveal information about their age or gender, and whether such cases are representative of the kinds of concerns their peer group generally has. These other questions could not be answered using just the original corpus, but the experiences of the initial study could be used to inform a later set of research questions for a follow-up study.

Indeed, one aspect of trying to answer the research questions that had been set by the NHS team was that the researchers realised that they were also answering a set of questions that they had not initially thought of and also had not been suggested by the NHS. A clue to these kinds of questions can be found if we look to feedback provided by two female patients, aged 20 and 83 years. In these cases, it was found that the patients were using aspects of their identity as a way of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

justifying or legitimating their position. As the analyses pertaining to the original 12 research questions were carried out, more cases were found where patients used language in various ways in order to represent themselves as worthy of attention or simply as being in the right. Three examples of these incidents are as follows:

> The nurse who saw me they were the rudest person I've ever come across in my life, they came out and shouted to the reception why I was I given the appointment as I was late they were shouting so I could hear, and when they called me in, they made me apologise I'm a married man with two kids.
>
> Been going there 12 yrs & rarely darken their doors. I am now left with what seems like placebo medication and disinclined to go back.
>
> My family also have been with [anonymised] Surgery ever since one was introduced, we have been a long standing and well respected family in the parish for over 400 years but, obviously this counts for nothing these days!

In the first example, the patient criticises a nurse who shouted to reception about him being late for his appointment, and then notes that he was made to apologise. At the end of his criticism, he provides his marital status and notes that he has two children. He doesn't elaborate on why this statement is relevant. A possible interpretation of the patient's self-description is that these aspects of his identity indicate that he is a 'grown-up' and thus he was unfairly treated like a child when made to apologise. Additionally, the statement might be interpreted as the patient implying that he should not be blamed for being late due to the responsibilities that come with fatherhood. In the second case, the patient constructs themself as a 'good patient' as opposed to someone who continually seeks medical help by mentioning that they 'rarely darken their doors'. Thus, they construct themself as someone who is able to give a credible opinion. Finally, the patient in the third example notes that they are from a family which has been respected locally for more than 400 years, a point which they relate with disappointment that it 'counts for nothing'.

It was found that patients regularly participated in this kind of legitimation work, and that there was evidence that different legitimation strategies were employed by different demographic groups. This aspect of the feedback was highlighted to the NHS, as it was reasoned that it is important to have awareness of these kinds of strategies, and to ensure that some forms of feedback are not privileged at the expense of others, just because some people are better at using certain strategies to strengthen the impact of their messages. Awareness about the potential power of legitimation strategies can therefore be important, as service providers need to make decisions about which kinds of feedback to respond to and in what ways.

The example relating to legitimation strategies was just one way that the researchers discovered they were answering questions that they had not

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

originally planned to address. As a result, after answering the 12 questions put forward by NHS England, the researchers added 4 new questions to the original list (see Baker et al., 2019):

13. Why do patients leave feedback?
14. What does the language of patients reveal about their expectations?
15. How do patients use language to construct their positions as legitimate?
16. What discourses do service providers draw on when responding to patient feedback?

A final aspect to mention about this study relates to timing. As previously noted, this project lasted 18 months and the researchers were given 12 questions to address. This is somewhat different from other corpus-assisted discourse analysis projects, where researchers are generally able to set their own questions or even to start the work with no questions. It was calculated that, taking holidays and weekends into account, the researchers would have about 28 working days to address each question by devising methods to analyse the data, then conducting the analysis and writing up the results. This was not a lot of time, and while they were able to provide reports for each question, it was felt that for some questions, more time would have been preferable in order to provide a more accurate or detailed set of responses. The NHS contacts were happy with the results that were given to them (so much so that they asked members of CASS to look at a second set of feedback, discussed in Chapter 10), and they did not make unreasonable demands relating to deadlines.

However, with other external partners there were perhaps somewhat unrealistic expectations regarding what members of CASS could produce within short timeframes (including expectations that they would work weekends in order to provide reports for meetings on Monday mornings). A key aspect of working with external partners, then, is to provide clear expectations about the amount of work that researchers are able to carry out within a given timeframe and the kinds of questions that can be answered easily (or not). A good organisation will be happy to view this kind of research as collaborative and as consisting of a dialogue where expectations on both sides may need to be adjusted occasionally. With that said, there was a definite benefit to having questions and deadlines set in advance by external partners, one which was perhaps not obvious from the outset but which became clearer as the project progressed.

2.5 Conclusion

Research questions are a central part of the process of doing research. As we have shown, however, different projects may involve different approaches to the development of research questions, with researchers potentially becoming involved at different points and being able to exercise different amounts of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

control on the nature of the questions. This applies particularly when doing research that crosses boundaries between disciplines and/or that involves interactions between researchers and practitioners, as in the case of our research on health communication. In addition, the power and flexibility of corpus linguistic methods make them potentially suitable and attractive for a wide variety of questions, data, and stakeholders, resulting in the different kinds of experiences we have presented.

Other things being equal, researchers may always wish for the freedom of exploration that we have described in relation to the study on media representation of obesity. However, we hope to have shown the value of considering the challenges and compromises involved in answering questions formulated more or less strictly by, in our case, people working in healthcare. Both the second and third case study resulted in findings relevant to practice and/or policy, they helped strengthen the relationships between CASS and valuable partners, and they forced CASS researchers to adapt and stretch their corpus linguistic expertise in ways that were interesting and more beneficial beyond each specific project.

The rest of this book will continue to show how doing corpus research on health communication often tested our ability to be flexible, adaptable, and creative not just at the start of the process but throughout. We will also continue to show how rewarding these experiences have been. In the next chapter, we turn to the topic of collecting data for the purposes of corpus construction.

References

Baker, P. and Brookes, G. (2022). *Analysing Language, Sex and Age in a Corpus of Patient Feedback: A Comparison of Approaches*. Cambridge University Press.

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

Baker, P., Gabrielatos, C. and McEnery, T. (2013). *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge University Press.

Bettini, M. (1995). *I classici nell'età dell'indiscrezione*. Einaudi.

Bordo, S. R. (1993). *Unbearable Weight: Feminism, Western Culture, and the Body*. University of California Press.

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

Dworkin, R. H., Turk, D. C., Revicki, D. A., Harding, G., Coyne, K. S., Peirce-Sander, S., Bhagwat, D., Everton, D., Burke, L. B., Cowan, P., Farrar, J. T., Hertz, S., Max, M. B., Rappaport, B. A. and Melzack, R. (2009). Development and Initial Validation of an Expanded and Revised Version of the Short-Form McGill Pain Questionnaire (SF-MPQ-2). *Pain*, *144*, 35–42. https://doi.org/10.1016/j.pain.2009.02.007.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002

Fernandez, E. and Boyle, G. J. (2002). Affective and Evaluative Descriptors of Pain in the McGill Pain Questionnaire: Reduction and Reorganization. *Journal of Pain*, *3* (1), 70–7. https://doi.org/10.1054/jpai.2001.xbcorr25530.

Gill, R., Henwood, K. and McLean, C. (2000). The Tyranny of the 'Six-Pack'? Culture in Psychology. In C. Squire (ed.), *Culture in Psychology* (pp. 100–17). Routledge.

Glaser, B. G. and Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine.

Gough, B. (2010). Promoting 'Masculinity' over Health: A Critical Analysis of Men's Health Promotion with Particular Reference to an Obesity Reduction Manual. In B. Gough and S. Robertson (eds.), *Men, Masculinities and Health: Critical Perspectives* (pp. 125–42). Palgrave Macmillan.

Gries, S. Th. (2011). Quantitative and Exploratory Corpus Approaches to Registers and Text Types. Plenary given at Corpus Linguistics 2011, University of Birmingham, 20–2, July 2011.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Routledge.

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.

Main, C. J. (2016). Pain Assessment in Context: A State of the Science Review of the McGill Pain Questionnaire 40 Years On. *Pain*, 157(7): 1387–99.

Marchi, A. (2014). A Corpus-Assisted Study of the Guardian's View on Journalism. Unpublished PhD thesis. Lancaster University.

Melzack R. (1975). The McGill Pain Questionnaire: Major Properties and Scoring Methods. *Pain*, *1*, 277–99. https://doi.org/10.1016/0304-3959(75)90044-5.

(1983). The McGill Pain Questionnaire. In R. Melzack (ed.), *Pain Measurement and Assessment* (pp. 41–7). Raven Press.

(1987). The Short-Form McGill Pain Questionnaire. *Pain*, *30*, 191–7. https://doi.org/10.1016/0304-3959(87)91074-8.

Scarry, E. (1987). *The Body in Pain: The Making and Unmaking of the World*. Oxford University Press.

Semino, E. (2010). Descriptions of Pain, Metaphor and Embodied Simulation. *Metaphor and Symbol*, *25*(4), 205–26. https://doi.org/10.1080/10926488.2010.510926.

Semino, E., Hardie, A. and Zakzrwska, 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.

Sunderland, J. (2010). Research Questions in Linguistics. In L. Litosseliti (ed.), *Research Methods in Linguistics* (pp. 9–28). Continuum.

Vadivelu, N., Urman, R. D. and Hines, R. L. (2011). *Essentials of Pain Management*. Springer.

Wilkie, D. G., Huan, H.-Y., Reilly, N. and Cain, K. C. (2001). Nociceptive and Neuropathic Pain in Patients with Lung Cancer: A Comparison of Pain Quality Descriptors. *Journal of Pain and Symptom Management*, *22*(5), 899–910. https://doi.org/10.1016/s0885-3924(01)00351-7.

Wray, A. and Bloomer, A. (2021). *Projects in Linguistics and Language Studies*, 3rd ed. Routledge.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 15:17:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.002