13 Conclusions

13.1 Introduction

In this final chapter of the book, we present a synthesis of the previous chapters. We first consider the question: what are the key insights into health communication that our different projects have given us? We then move on to consider the lessons we learned about carrying out corpus-based research on health communication, offering practical advice and tips relating to research questions, datasets, analytical approaches, and going beyond academia. This is followed by a section which critically considers the limitations of the corpus-based approach. And, finally, we consider future directions for corpus-assisted healthcare research, asking what has changed since we completed the projects described in this book and what avenues of research we believe are potentially interesting to investigate next.

13.2 What Have We Learnt about Health Communication That We Did Not Know Before?

In this section, rather than reiterating some of the main findings from the parts of individual studies described in earlier chapters, we instead want to focus on some of our 'bigger picture' findings, which tend to stretch across and connect with multiple projects.

First, health communication is not restricted to health practitioners or even communication between health practitioners and patients. In particular, people who experience health conditions are not passive – they co-construct understandings around their conditions on their own terms, without their medical practitioners. The data we examined was incredibly *human*. We found human nature displayed in comedic or cute ways when finding metaphors to frame health conditions – from the long-running humorous reframing of cancer patients as members of an army to the creation of a cartoon-like 'Mr Anxiety'. We found that human nature could sometimes be articulated through fear or prejudice – from concerns that a vaccination might kill you to the characterisation of syphilis as a 'French' disease. And there were even aspects

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

of human nature that manifested as entitled or petty – from a patient's complaint that they expected better treatment because their family had lived in the same area for hundreds of years through to news articles gloating about murderers putting on weight in prison. For some of us, prior to working with corpora of health data, we had expected that we would be analysing a very dry, scientific form of discourse. This was rarely the case. Sometimes the texts we read could be extremely funny, frustrating, or moving. The data was thus much more engaging than expected, although that could also bring challenges with it, especially when the material was potentially distressing and we were aiming for as much objectivity as possible in our approach to it.

Collectively, the corpora collected for our projects show how health communication extends across a much wider range of linguistic 'events' than, say, an appointment with a General Practitioner. People gain understandings about health from a wide variety of sources: friends and family, online forums, governments, scientific researchers and the media, and these sources interact with one another – nobody can be said to be truly impartial or beyond influence from the discourses of others. Although health conditions are real and exist beyond discourse (as cancer and COVID-19 can kill people), the ways that we understand and react to them are dependent on discourse, and language plays a major role in conveying, challenging, and upholding these different understandings. Language is where we co-construct beliefs about what counts as a health condition and what counts as healthy. And there is a lot more variation than we had expected to find – for example, when we observed that Violence metaphors for cancer seem to be empowering for some patients. Fortunately, a corpus-based approach is well-suited to explore and identify a lot of this variation. With millions of words of naturally occurring data, we were able to confidently make generalisations about trends in language use, while also spotting the less frequent patterns which may have been missing from smaller datasets.

Similarly, the ways that we use language to communicate about health are varied – in some of our projects we took a prospective view, allowing different forms of language to emerge during our analyses. We had not expected there to be so many metaphors across different projects, but there was also abundant use of transitivity, evaluation, legitimation, narrative, humour, punning, emojis, alliteration, and anthropomorphisation. The authors of this book all have backgrounds in linguistics, which proved to be helpful in identifying the wide range of phenomena encountered throughout the different corpora, but even we were surprised by the extent of linguistic variation across each project. With each new corpus, we had to set aside what we had done and start again, with fresh eyes.

One aspect of our research that we had not expected to play such an important role was identity. A lot the variation that we found can be accounted

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

for through identity variables – it predicted differences in language use in patient feedback and in the ways that people on online forums framed their health conditions. It also played a key role in the ways that journalists wrote about health – obesity is *very* gendered in the news. We all hold multiple identities, which can shift in and out of focus in different contexts, like kaleidoscope patterns. Some identity characteristics can be easier to identify and compare than others, though; the challenge for analysts is to consider which ones are most relevant and which are missing but ought to be interrogated. We also need to consider which identities interact together (sex *and* age helped explain variation in patient feedback relating to cancer, for example). Another key factor in terms of variation is time, and our analyses have shown how the consideration of time can go beyond merely dividing a corpus up into years based on date of publication of texts but can also involve annual patterns, the age of the contributor, or the length of time they have spent in a particular discourse community.

Health communication research can perhaps be characterised as action-oriented, in that it aims to improve understandings of language use around health in order to foster better health outcomes for people. Many researchers in this field (particularly the subfield of corpus-based health research) tend to take a descriptive rather than an evaluative view. Often, we do not know what we are going to find in a corpus, and so we generally do not set out to 'prove a point' or be critical of people's language use. So, in our analyses of metaphors around cancer, we created a metaphor menu as opposed to suggesting that some metaphors were bad, while in our analyses of an anxiety forum, we were cautious about suggesting that some framings of anxiety were harmful. On the other hand, in some of our projects, we have tried to offer more direct advice – such as making suggestions to improve the descriptors used in the pain questionnaire, consider how the NHS could use information campaigns to change patient expectations, or determine how journalists could write about people with obesity or dementia in ways that are less stigmatising. The answer to the question of what to do with a finding also varied tremendously across our projects. While Chapter 2 outlines the ways we created research questions under different conditions, a point to bear in mind is that all of these projects had a similar overall goal – how to do the most good. This was not a question we explicitly considered when we were focussed on each project, but in hindsight, we realise it is the most important one.

13.3 What Advice Would We Pass on to Other Corpus Researchers Working in Health Communication?

When you read an account of a research project in an academic journal, book, or newspaper report, it has usually been tidied up – with false starts, backtracks,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

dead ends, and loose endings all made magically invisible, as if they never happened. In other words, this is often a simplified version of what actually happened. Projects are rarely like that, though. They can be messy and even go horribly wrong in both foreseen and unexpected ways. In some of the worst cases, they can fail to produce anything of value. The projects we described in this book *were* successful, although they did not always go as planned. So, with the accumulated knowledge of all these projects, what tips would we give to other corpus researchers in health communication? What do we wish we had known from the start?

First, corpus research in health communication is often best achieved through teamwork. The approach requires quite a wide-ranging skill set, and it is unlikely that a single person will be able to tick every box on the list. Computational knowledge is useful for building, cleaning, and annotating corpora, then mounting it on analysis software. Statistical skills are required to make sense of what the tests are doing, which ones should be used, and what the settings should be. Linguistic skills are needed in order to identify and interpret the features in a corpus. Depending on the corpus under examination, we may also need a specialist historian or someone with detailed knowledge about a particular social or political context. And it is also very important to involve someone with knowledge about the particular health condition or healthcare setting. All of the corpus-based projects we describe in this book involved more than one person, and frequently, they involved someone who was *not* a corpus linguist. It perhaps seems counter-intuitive to say that a corpus project should actively seek to recruit someone who is not knowledgeable about corpus linguistics, but for health-related communication, there are advantages to be gained. The non–corpus linguist can provide a better sense of what matters in the health-based context under examination. They can give the research a clearer and more relevant focus, and help interpret and explain results. As we saw in Chapter 2, they can also push corpus linguists out of their analytical comfort zone, ensuring that they are not stuck doing repetitive 'handle-turning' forms of research. Thus, we would advise viewing this kind of research as a continuing dialogue between multiple participants with different areas of expertise. However, as Chapter 12 showed, relationships between those working in academia and those connected to health organisations do not always go as planned. Therefore, the dialogue should also involve a focus on scheduling and dividing tasks, allowing the different parties to compare their organisational structures in order to set expectations and boundaries. This might save time and avoid disappointment at a later date.

In terms of creating a corpus, it is important to consider what any initial research questions are and not to spend so long creating the perfect corpus that there is a suboptimal amount of time required to analyse it – even if this means collecting less data or tolerating a certain level of messiness in the corpus itself,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

as with the OCR errors in the VicVaDis corpus. Unless you are certain that a hand-annotated corpus is required, this is probably not something to embark on at the outset; you can also add in those annotations later if they become essential. Some form of automatic tagging can be a more pragmatic option initially, although we would advise analysts to be mistrustful of the accuracy of automatic tags; for example, when working with a grammatical tagged version of the news corpus on obesity, we found numerous cases of the word *fat* tagged as an adjective when it should have been a noun. Such cases had to be weeded out by hand, and if we had just taken the cases at face value, we would have achieved very different (and much less accurate) results. Expect that there will be both anticipated and unanticipated tagging errors and keep an eye out for them, making adjustments to your calculations if and when needed.

A pilot analysis can be useful in terms of helping spot potential problems with a corpus. Even when we worked with corpora that had not been tagged, we discovered numerous instances of duplicated files or unwanted boilerplates (such as repeated menu headings from websites or copyright information) which skewed frequencies and gave us keywords and collocates that were less accurate or useful than they should have been. Be prepared *not* to trust your text, in other words, and to view the initial analyses more as trouble-shooting exercises, aimed at weeding out these kinds of problems with data collection.

Some forms of corpora bring with them their own challenges. Spoken corpora are especially time-consuming to collect and transcribe, in addition to posing some of the most complex ethical challenges involving anonymisation and consent. When working with most health-based topics, it is important to take ethics into account (with exceptions such as texts widely intended for public consumption like newspaper articles), and this may mean that there are some forms of data that simply can't be collected or that we can't have full knowledge of. For example, with the forum posts on pain and anxiety, some posters had not consented to their data being used for research, so their posts had been removed in advance of us receiving the data. When working with data of an interactive nature, this reduces the kinds of analysis we can confidently carry out, and sometimes we have to accept that what we have is not ideal – although it can still tell us other things. Ethics aside, one concerning aspect that we noted while we worked on these projects is that some forms of data are becoming more difficult to obtain; X (formerly Twitter) has placed restrictions on how its social media posts can be collected, while the LexisNexis online news aggregator has undergone several changes to its database in recent years. At one point, the site required users to manually tick a check box for each article they wanted to collect. More recently, the database has limited the number of articles that can be collected in one day to 1,000. Owners of online data are understandably concerned about mass scraping of their data, which has

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

sometimes been used without permission in large language models for AI chatbots. The larger point we want to make is that data collection protocols can change rapidly. Don't assume, from reading about one of the studies in this book, that your experience of collecting data will be the same.

For our projects involving corpora of speech, such as Emergency Departments conversations or interviews relating to voice-hearing, we had to rely solely on human transcribers – although more recently new technologies have helped improve transcription, both in terms of speed and cost. For example, Sonix[1] is an online audio and video transcription software which we have since used in other corpus projects. The tool does not provide 100 per cent accuracy, but it greatly reduces the amount of work that human transcribers need to do. Mobile phone technology has become much more impressive over the past decade – point a phone at a page of printed text and its camera can likely scan and create an electronic version of it. Even handwritten pages can be converted in this way, although they tend to be less accurately rendered. When reading about how others created corpora, don't assume that the technology has stood still and you should replicate older methods of data collection. The same point applies to corpus analysis tools; during the period in which we worked on these projects, newer versions of existing tools were launched, enabling a wider range of forms of analysis. While it is important, then, to refer to existing literature, this is a field which is moving quickly – a warning that by the time any piece of corpus research has been published, it is already out of date.

Readers of this book will have noticed that we did not rely on a single corpus tool for all our projects, but we used a variety of them: Sketch Engine, AntConc, CQPweb, WordSmith, and Wmatrix. In some cases, choice of software was governed by our existing familiarity with certain tools; in others, it was influenced by the different affordances that tools allow. Sometimes multiple tools needed to be used with the same corpus, although it should be borne in mind that slightly different results can be obtained. For example, different tool creators might have their own views on how to define a 'token' (some may advocate splitting hyphenated words into two tokens, some may not), and this can impact on frequency counts. In addition, some tools may have different means of calculating collocation or keyness. Aim for consistency, if possible, and be clear about which tool was used for each procedure.

The different chapters across this book have also indicated that there is no single approach or pathway to carrying out analysis. Instead, we are given a bunch of analytical techniques and can choose to apply some or all of them in different orders, using different cut-off points or tests for statistical significance. The lack of a single route can be challenging but also liberating – enabling

[1] https://sonix.ai/.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

a more experimental, ludic approach to analysing a corpus. It can be helpful, at least at first, to try to take into account the goals of the project as well as the nature of the corpus. Sometimes we might want to focus on a single word or phrase in the corpus which is considered to be important (such as the word *anxiety* or *pain*). In other cases, we may want to look at the corpus as a whole. Sometimes we may decide in advance which kinds of linguistic features we want to examine – such as metaphors – while in other cases we might want to allow salient or frequent linguistic features to emerge. There are advantages and limitations to both approaches, and a certain amount of reflective modesty is advised when outlining findings and implications. Of course, there is no reason why a combination of approaches cannot be taken, and sometimes shifting between a mixture of targeted and prospective techniques can work well.

Another piece of advice we would give is to be prepared to change course, to allow the corpus analysis to reveal new and unexpected avenues. Be on the lookout for answers to questions that you did not think of but are actually more interesting than the ones you originally asked. For example, in the study on venereal disease discussed in Chapter 8, analysis of collocates revealed so many place names that this shifted the nature of the research. And in the study on patient feedback from Chapter 6, the analysts gradually realised that the nature of the data involved a lot of legitimation alongside the evaluation – legitimation that lent itself to questions that neither the corpus linguists nor the NHS team had originally thought of.

Another piece of advice we can't stress enough is to avoid making assumptions about decontextualised language use. You are likely to be looking at a lot of lists of words and phrases with numbers alongside them. We might be tempted to guess at what these words mean or imply for our data, but experience has shown us again and again that we can often be completely wrong. This is a case where we need to trust the *context* – reading concordance lines, sometimes expanded concordance lines, in order to truly get a sense of how a linguistic item is being used. It is worth trying to achieve a balance between covering lots of linguistic items while also getting a reasonably accurate account of them. If a word occurs 20,000 times in your corpus, we would advise that it is not a good use of your time (and sanity) to look through all 20,000 lines. In many cases, examining a random selection of 100 lines is likely to identify the main trends. If this provides inconclusive results, the selection can be expanded. A thousand lines is likely to give a good selection of rarer cases (although not all the rare cases).

Corpus research is especially well-suited to impact the world outside of academia. The sheer size of our datasets enables our findings to be taken seriously, while our combination of quantitative and qualitative approaches means that we often have opportunities to make unexpected insights into language use. The challenge, then, is in conveying our findings to those

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

outside academia – particularly as this is an area where we may not have as much experience or training. Be mindful that an academic style is unlikely to be appropriate for most forms of impact. In this book we have written about unlearning academic writing skills and thinking about specific audiences (in some cases very specific – say a General Practitioner reading while eating a sandwich for lunch). Rather than crossing one's fingers and hoping that our attempts to communicate outside our domain will work, it is worth spending some time reading existing news articles and press releases that discuss health-related or corpus-based academic research, while critically engaging with them to get a sense of what works. Non-academic members of a research team can be especially valuable in helping frame these kinds of dissemination texts, while also ensuring that the style is appropriate. Short, unambiguous, surprising messages tend to work well, so be sure to get advice if this style of writing is not your forte and consider some media training if you are going to be giving interviews as well. Also, heed a warning from Chapter 12: it is worth considering the ways that your message might get mangled. If possible, ask a journalist to send a copy of their news story back to you, prior to publication (and raise a red flag if they are reluctant to do so).

13.4 What Are the Limitations of the Corpus-Based Approach?

No approach can do everything well, and we hope that this book has given readers a sense of where the corpus approach can shine and where it can only take us so far.

One limitation relates to the kinds of data that are available to us in large enough amounts for a corpus approach to be considered. While even a small data set can be referred to as a corpus, a few thousand words is likely to be short enough for a qualitative close reading to be carried out on it, and unless the corpus contains a lot of lexical repetition, frequencies are unlikely to be high enough for much of interest to emerge through techniques like keywords or collocation.

Some texts are easier to collect than others, which can push analysis into certain directions and away from others. Written data is usually easier to source than spoken data. Recent data is easier than historic data (as a general rule, the further back in time you go, the harder it is to build a corpus). Spoken historical data can be extremely difficult to find. Online data can be an easier option, then, although even here there can be potential access difficulties, and the presence of bots can sometimes make us question the veracity of such data. Note that large amounts of text can bring their own problems. Some analysis tools will crash or work very slowly when working with millions or billions of words of data. And the more text we have, the harder it is to fully understand, risking errors of interpretation.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

The bedrock of the corpus approach is frequency, and that can reveal a great many insights, although not all. Some things are more difficult to identify and count than others; hence there can be a risk that we limit our analyses to simple word frequencies, as opposed to, say, more complex and variable phenomena like metaphor or joking.

Two corpora may have similar relative frequencies of a word, but in each corpus that word might be used very differently – something which we may miss if we only carry out a keywords analysis. And techniques like keywords prompt us to consider frequency differences as being important; they often are, but it might be the case that the *similarities* between two corpora are also relevant. Frequency can also make us focus on presence – we may spend so long counting what is in a corpus that we don't notice what isn't there. So think about the *absence* of linguistic features as well as presence. What could be there, or should be there, but isn't? It is thus worth reflecting on how the analytical methods you employ might be limiting you or steering you in certain directions. Comparing frequencies across multiple corpora might help identify the complete absence of a feature in one of them, but what if the feature occurs in none of the corpora?

We advise researchers to consider whether a corpus approach is actually going to enable you to answer your research questions effectively. In some of the studies we outlined in this book, we concluded that using tools like collocation and keywords alone would not get us far, so we switched to a more qualitative, manual analysis (e.g., the annotation of discourse functions in the anxiety forum corpus, as well as the identification of metaphor in the MELC corpus). While such approaches can be time-consuming, they don't necessarily need to involve the entire corpus – sometimes a downsampled set will provide enough evidence for us to be able to spot trends or carry out comparisons of the most frequent features, even if we can't say that the analysis will be exhaustive.

The techniques of corpus linguistics are descriptive. They can tell us what is happening with language, but they can't tell us what this means, or why it is happening, or whether this should be happening. In contexts like health communication, the corpus approach needs supplementing with consideration of relevant forms of context so we can interpret, explain, and critique the findings. So, for example, the analysis of the newspaper corpus of articles about obesity revealed the ways that journalists write negatively about people with obesity, using shaming and ridiculing language. However, the corpus analysis alone can't tell us why journalists did that. We would need to think more about the context of news reporting in the country and time period under study, along with taking into account aspects like government policy, press regulation, readership demographics, and vested economic concerns. In interpreting and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

evaluating our findings, we might want to consult with stakeholder groups who are likely to be able to offer real-world insights based on lived experience.

13.5 Final Thoughts: What about the Future?

The projects described in this book ran between the years 2012 and 2023, with the most recent corpus texts being in data from Mumsnet, which includes contributions posted up to the end of 2022. What changes have taken place since that period that are relevant for health communication? In the UK there have been unprecedented decreases in satisfaction with the NHS – from an average 53 per cent satisfaction in 2020 to 24 per cent in 2023.[2] Our two patient feedback studies covered earlier time periods, where we found a mostly positive picture. The change indicates how time-limited such research can be, as well as the need to continue analysing feedback data in order to respond to contexts that are in constant flux.

Linked to changes in patient satisfaction, the political context has also changed. For our study on press language around obesity, we collected news articles published between 2008 and 2017. The majority of the articles in that corpus, then, had been written under a Conservative-led government in the UK. Perhaps unsurprisingly, we found that news framings about obesity tended to be largely congruent with the dominant political ideologies of the time, for example, by stressing personal contexts while reducing the role of larger social structures. Consequently, there had been a move away from discussing obesity in terms of issues like inequality and poverty, despite the fact that such phenomena had increased over the period under study. In 2024, a Labour government came to power; shifts in terms of the ideologies and policies of those who run a country tend to impact on health policy, and it will be interesting to see how a new set of leaders will influence language use relating to health, not just in the press but in a wide range of communicational contexts.

There have also been changes in terms of a biomedical perspective. Sticking with obesity – in 2022, a review of anti-obesity treatments concluded that semaglutide (an antidiabetic medication) was more promising than previous anti-obesity drugs, and in 2023, a brand of the drug called Wegovy was approved for use by the NHS for weight loss. Demand for such drugs appears to be rising, popularised by celebrity endorsements – particularly in the US. *The New York Times* published an opinion piece on 23 October 2023 which claimed ‘these drugs are blockbusters because they promise to solve a medical problem that is also a cultural problem – how to cure the moral crisis of fat

[2] https://natcen.ac.uk/publications/public-attitudes-nhs-and-social-care#:~:text=In%202023%2C%20fewer%20than%201,public%20satisfaction%20with%20the%20NHS.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

bodies that refuse to get and stay thin'.[3] It isn't difficult to see how scientific advances can completely change the discourses around health conditions. These examples show how research which looks at health communication in contemporary contexts needs to be ongoing in order to keep up with relevant developments in a quickly changing world. As noted earlier, frustratingly, by the time that a corpus has been created and analysed, and the research published, it can already feel slightly out of date. This is less the case for research which considers historical contexts, like the studies on venereal disease and vaccine hesitancy discussed in Chapter 8, and it is interesting to see how centuries-old texts can still shed light on the present day. Consequently, corpus studies using recent data, while situating the findings within that context, should also articulate more solid and lasting findings and implications from their projects.

Artificial intelligence and systems like ChatGPT appear, on face value, to be able to answer any question or analyse any text which is presented to them. These systems were not available when we carried out our corpus research, which involved a lot of human-led decision-making, analysis, and interpretation. Since then, ChatGPT has been incorporated into AntConc, and some of us have experimented with the potential for AI tools to aid in corpus analysis (see Curry et al., 2024), finding a mixed picture (i.e., not one in which we believe AI could replace human researchers, at least not at this point in time). ChatGPT did a reasonable job of putting keywords into thematic categories but had difficulty in interpreting concordance lines where knowledge of context was required. Broadly, it could produce a piece of corpus research which might achieve a low pass mark if submitted as an undergraduate essay but would not be publishable.

On the whole, though, computational advances present opportunities for corpus researchers. Often the texts that are taken to be compiled into a corpus are a mixture of writing and images, with the latter elements usually stripped out (such was the case for the obesity news corpus). However, AI tools are becoming adept at tagging images with labels, based on millions of cases of existing pre-tagged data. We have experimented with one tool (Vertex AI, formerly called Google Cloud Vision), working with a small corpus of news articles about obesity (Baker and Collins, 2023). Incorporating image tags into the corpus allows the analysis to be truly multimodal. For example, we were able to consider which newspapers tended to use which types of images and the extent to which particular words appeared in articles that contained certain images. Our analysis helped us show how stories that sympathetically focussed on people's struggles with obesity tended to show them in formalwear at public events, whereas those which focussed on body positivity were more likely to

[3] www.nytimes.com/2023/10/09/opinion/ozempic-obesity-fat-diabetes.html.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

have pictures of women in revealing clothing. There is a great deal of potential, then, for corpus analyses to consider the relationship between words and image.

It is challenging to try to identify future topics, even when extrapolating from current trends. For example, improved healthcare across the globe is helping to increase lifespan, which is likely to have implications for the kinds of health care that will be needed and talked about. In recent years, there have been greater numbers of people reporting mental health problems over time, as well as more people diagnosed with forms of neurodiversity. Both indicate future avenues of health communication research to be explored. Additionally, rising average global temperatures due to humanity's burning of fossil fuels are likely to result in increases in a range of different health conditions, such as Lyme disease, West Nile virus, cardiovascular disease, allergies, and asthma, in addition to impacts on mental health. Another pandemic is possible in the not-too-distant future, while advances in immunotherapy for cancer are beginning to involve personalised 'vaccines' that train the immune system to recognise and kill the patient's cancer cells. Increased automatisation of healthcare through AI and robotics is also likely to suggest new directions for analysis. For example, an offshoot of our corpus analysis of the anxiety forum involved us collaborating with the charity Anxiety UK in order to analyse human interactions with a chatbot on its website; corpus analysis helped us discuss with the charity how the chatbot's responses could be improved (Collins et al., 2024).

Finally, the research outlined in this book was carried out at a British university with funding from UK funders and mainly had a British focus, with some exceptions (the Emergency Departments corpus contained Australian data, while the health and pain forums contained significant amounts of posts written by American authors). On the whole, however, we have focussed on the UK and the English language – taking advantage of our familiarity with this context, which helped in terms of providing interpretations and explanations of our findings. We want to make it clear, though, that the techniques described in this book can be used with all languages, and we would hope to encourage future researchers to carry out corpus-based analyses of health communication in an ever-broadening range of geopolitical and historical contexts. There is also much to be gained from research which takes comparable datasets from different countries and time periods, identifying differences and similarities and basing interpretations on different social, political, and economic contexts and how they intersect with understandings of health. It should not be assumed that such projects will be straightforward in terms of locating and gaining access to data. The political systems in some countries may make it harder to gain access to certain kinds of health-related texts. Similarly, in some countries it can be difficult to publish anything which might be interpreted as being critical of the government, with resulting

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013

implications for objectivity. The Global Expression Report (2023) found that only 13 per cent of the world's population live in 'open' countries, while 34 per cent of people live in countries where freedom of expression is in crisis. There needs to be more thought and effort to enable health communication research to represent the health experiences and concerns of the whole world.

In closing, we hope that this collection has helped provide a sense of the scope and techniques associated with a corpus-based approach to health communication. We also hope that we have conveyed the value and importance of this approach. Our aim in writing this book has been to encourage and inspire others to work in this field – all the studies that we describe here were both challenging and rewarding, and we feel that we have learnt a lot from our experiences and grown as researchers. Our team started from a relative position of ignorance, with a steep learning curve. Our goal is to hopefully make that learning curve a little less daunting for others.

References

Baker, P. and Collins, L. (2023). Creating and Analysing a Multimodal Corpus of News Texts with Google Cloud Vision's Automatic Image Tagger. *Applied Corpus Linguistics*, *3*(1), 100043. https://doi.org/10.1016/j.acorp.2023.100043.

Collins, L., Nicholson, N., Lidbetter, N., Smithson, D. and Baker, P. (2024). Implementation of Anxiety UK's Ask Anxia® Chatbot Service: Lessons Learned. *JMIR Human Factors*, *11*, e53897. https://doi.org/10.2196/53897.

Curry, N., Baker, P. and Brookes, G. (2024). Generative AI for Corpus Approaches to Discourse Studies: A Critical Evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), 10082. https://doi.org/10.1016/j.acorp.2023.100082.

Global Express Report. (2023). *Article 19*. www.article19.org/resources/the-global-expression-report-2023/?gad_source=1&gclid=Cj0KCQjw0ruyBhDuARIsANSZ3wobAv2bd1ZAjE41VoGcRox-BWpiXfmET7CO1NY6uoJ_bLsOAVH4tRsaAn2vEALw_wcB.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 07:14:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009477680.013