McEnery, Anthony, and Helen Baker. "Introduction." *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History*. London: Bloomsbury Academic, 2017. 1–31. Corpus and Discourse. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781474295062.0005>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 09:15 UTC.

Copyright © Anthony McEnery and Helen Baker 2017. Released under a CC BY-NC-ND licence (<https://creativecommons.org/licenses/by-nc-nd/3.0/>). You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

Introduction

1.1 The aim of this book

This is a book about using corpora – large bodies of machine-readable texts – to assist historians to gain insights into people, societies and cultures long gone. In this book we do not want to focus on a narrow question and pursue it. We have chosen precisely to draw on a large canvas specifically to test the limits of – and the opportunities presented by – the corpus approach to the study of history. In looking at this we want to work, as much is possible, in collaboration – we do not want to see the subject expertise of the historian as an adornment to the work of a linguist. Nor do we wish to see a linguist as an expert who can be called upon to examine a narrow linguistic issue and then be dismissed. We want to explore, in a spirit of full cooperation, what a linguist and historian can achieve together. We think that using the field of corpus linguistics as a setting in which that collaboration occurs can be especially fruitful. Both corpus linguists and historians have an interest in approaching and assessing large bodies of, typically documentary, evidence. Similarly, both also understand and value the benefits that close readings of individual documents can bring to the study of a much larger text collection. So the methodology of corpus linguistics creates, in our opinion, a common ground on which the linguist and historian can meet. The linguist brings expertise in the manipulation and understanding of large textual databases, with experience of the insights that can be brought about by the appropriate use of the tools of the corpus linguist. The historian brings their knowledge of the period and topic in question, with a well-developed sense of the hypotheses of interest to that subject community and a body of work that can help to frame and explain what the corpus investigation finds. Both also approach the collaboration wishing to be challenged – what historical approaches set methodological and analytical challenges for the linguist? What findings from the corpus present challenges for established historical research? Such questions

underlie the enterprise that led to this book, a book which welcomes a healthy, bidirectional trade in ideas and investigation between linguistics and history.

But what subject have we chosen to explore to realize this collaboration and why? We have chosen to study women who traded sex for cash or some benefit in kind in the seventeenth century. For ease of reference we will call that prostitution, though our findings, as will be apparent as the book progresses, fall at times outside of what we might think of as prostitution today. Seventeenth-century prostitutes, usually coming from the ranks of the lower classes, were mostly illiterate and thus rarely speak to us directly through recorded history. However, we are setting ourselves the challenge of exploring *general* attitudes towards prostitution in what was a turbulent century. We are interested in particular in seeing whether, in looking at general written English from the period, we are able to gain insights into the representation of prostitution in the period – how people wrote and thought about the people involved in the sex trade and the trade itself. Given a span of hundred years, we are interested to see how, if at all, that representation changed over the century. Where it does, we will be interested to seek to explain how that change came about supporting, where we can, our hypotheses with historical research and corpus analyses.

A goal we set ourselves from the outset of our work was to look at public discourse – general written English of the period. Studying specialized documents such as court records, for example, would doubtless have yielded much discussion of the sex trade. However, this is not what we are interested in. We want to look at general attitudes to prostitution – not judicial attitudes to it. Similarly, where we study texts, we are interested in seeing what is said about prostitution outside of the confines of a courtroom where it was being judged and condemned. When away from the court and other sources of judicial power, what was said about transactional sex and those involved in it in the seventeenth century? In exploring such a question we are, of course, aware that censorship regimes in the century may be an explanation for some of what we see – and what we do not see.¹ We are also mindful that censorship in the period can be overt (e.g. legislative) and covert (e.g. the judgement exercised by the owner of a press over what they were and were not prepared to print). Nonetheless, we think that the question of the everyday and the opportunity that this brings to unearth the views of those with less overt power than is present in a courtroom makes this investigation worthwhile.

The investigation is also worthwhile from the point of view of a corpus linguist. The challenge set by this question is substantial. As will be explored in the remainder of this chapter, the challenge is to work with corpora the scale

of which has not previously been used to explore language in the seventeenth century. Given the scale of the corpora involved and the breadth of the historical investigation undertaken, the research also makes new methodological demands that this book will explore. However, for both the historian and linguist we believe the effort is worth it. The spur to innovation that working with a new subject brings makes demands on, and expands the horizons of, corpus linguistics. The scale of investigation that corpus linguistics can, in principle, offer to an historian makes a venture such as the one reported on in this book worthwhile – to be able to account for millions of words of data across a long time span and to test hypotheses from small-scale, close reading, against large-scale, machine-aided analysis holds real promise.

However, it should be apparent from what has been written so far that the task set by the historian for the linguist in this book is not trivial. What corpus can we use to explore this topic in this century? What tools should be used? In this chapter we will explore these questions. To begin, let us consider the current state of the art in historical corpus linguistics and how it needs to progress, in order to enable exploration of a broader range of research questions of relevance to linguists and historians.

1.2 Looking into the past

In spite of a large body of work in linguistics in general, and in corpus linguistics in particular, using corpora to explore the past is still in its infancy. Pioneering work with corpora such as the Helsinki corpus,² the Archer corpus³ and the Corpus of Early English Correspondence⁴ has provided a wealth of useful insights into the English language in the past. Even in the area of speech, the work of Culpeper and Kytö (2010) has allowed a window onto the spoken language in the early modern period. By collecting representations of speech in sources such as court records and plays, Culpeper and Kytö devised an ingenious method for at least beginning to look at spoken English in a period well before the creation of sound recording devices. Books have been written looking at the development of English over time (Leech et al. 2009) and whole journals are devoted to looking at features of interaction in language in the past, such as the *Journal of Historical Pragmatics* and the *Journal of Historical Sociolinguistics*. Such a host of corpora and publications using corpus linguistics to investigate the past of English hardly seems to give credence to our claim that the use of corpora to explore the past is in its infancy. Yet it is.

The first issue that we encounter when we look at the corpora that have been used to explore the past is that they are typically small. Take the Archer corpus, for example. It is a corpus of 3,298,080 words looking at a number of discrete time periods of fifty years in length over a period from 1600 to 1959 for British English and from 1750 to 1959 for American English.⁵ Similarly, the Helsinki corpus is small. It covers a period of 950 years, from 750 to 1700 AD, split into Old English (with sections covering 750–850, 850–950, 950–1050 and 1050–1150), Middle English (with sections covering 1150–1250, 1250–1350, 1350–1420 and 1420–1500) and Early Modern English (with sections covering 1500–70, 1570–1640 and 1640–1710), yet it does so using only 1,572,600 words.⁶ Neither the Archer nor the Helsinki corpus is unusual in using such a relatively small size to cover a comparatively large span of time. In fairness, the size and structure of those corpora reflect very much the types of research questions the corpora were designed to address. Both corpora have been used to look at the development of the grammar of the language over a long period of time. The grammatical features of a language often change very slowly, if at all. Many of the features being looked at, such as modals,⁷ are very frequent. So a relatively small corpus composed of a series of ‘snapshots’ in which the language at a point in time is sampled, is quite well designed to explore frequent grammatical features changing over time. Similarly, the Corpus of English Dialogues also permits a number of interesting and frequent features, such as discourse markers (Lutzky 2012), to be explored through what is, in reality, a corpus of relatively modest size. However, the issue of size becomes acute when one tries to use such corpora to explore words of what one might describe as moderate or low frequency; there is simply not enough data to begin to make generalizations of any worth. Take, for instance, the word *prostitute*. Table 1.1 shows how many examples of the use of the word occur in the data from the seventeenth century in three corpora.

Table 1.1 Number of examples of the word *prostitute* in three well-known historical corpora

Corpus	Period	Number of words covered by that period in the corpus	Examples of <i>prostitute</i>	Normalized frequency to per 100,000 words
<i>Archer</i>	1600–99	397,869	2	0.502
<i>Corpus of English Dialogues</i>	1600–1719	924,765	0	0
<i>Helsinki Corpus</i>	1570–1710	416,617	0	0

It may be that the word *prostitute* is infrequent in Early Modern English. On the other hand, it may be that it is moderately frequent, but the size of the corpora means that they do not yield many examples because some simply miss having any examples of the word (Corpus of English Dialogues, Helsinki Corpus). Another has but two examples of the word (Archer). Perhaps, one may conclude, the word was very rare in the seventeenth century. On the basis of what is, in fact, very little data, it would be unwise to jump to that conclusion. The point is well illustrated by considering the frequency per 100,000 figures in Table 1.1 and comparing those to what we find in corpora of present-day English. The British National Corpus (Aston and Burnard, 1998) contains approximately 88 million words of written English, principally from the early 1990s. In this corpus, the word *prostitute* occurs 307 times, or 0.349 times per 100,000 words. The British English 2006 corpus (Baker 2009) contains approximately a million words of written British English from around 2006. It contains four examples of *prostitute*, or 0.348 times per 100,000. The word seems no more rare in these corpora than in the Archer corpus. No speaker of present-day English would think of *prostitute* as a rare or especially infrequent word. So the suspicion remains that, with the small corpora of seventeenth-century English, we are simply missing the examples to investigate words which were, in fact, in quite wide circulation and moderately frequently used because the corpora available are too small.

An analogy may help here. Imagine that you are in a room with a million boxes. Each contains a picture of an animal. You are asked to estimate how many contain pictures of rabbits. You do not have time to look in all of the boxes so you look quickly in a hundred. You find no pictures of rabbits. You do, however, find five pictures of hamsters in the boxes and one of a fox. You conclude that in the population of pictures in the boxes, ones of rabbits are at best rare, those of hamsters are very common and the ones of foxes seem more common than rabbits. Another person, with more time, looks at 100,000 boxes. They confirm that pictures of hamsters are common – they found 4,833. Yet they also found 500 pictures of rabbits, 480 of horses and 1 of a fox. They conclude that you looked in too few boxes to get a fair idea of what was in the population of boxes. You missed a moderately frequent picture as a result – that of the rabbit – and failed completely to notice that there were pictures other than those of hamsters and rabbits to be found. You also spotted a very rare feature, the fox, but misclassified it as more frequent than the rabbit as chance led you to find the fox but not a rabbit, and your sampling of the boxes was too small to correct that error.

What is the purpose of this analogy? The point is that when you are looking at a very large and diverse feature – in our case words – the size of your sample needs to be proportionate to the frequency of the feature you are looking for. The corpora we have used to date are like the person who searched the boxes first. That search allows you to spot the most frequent features with some reliability. The second person looked at a much larger proportion of the boxes and hence was able to identify what one may call moderately frequent features – the rabbits and horses. They were also able to identify what looked like a truly infrequent feature. So the corpora we have used to date to explore the past have allowed us to identify the frequent features. But they are too small to allow us to begin to reliably discriminate moderately rare features from moderately frequent features. They also give us too few examples of moderately frequent features to be of help.

When we are investigating content words – where our focus is on meaning rather than grammar, as it is in this book – we need larger corpora than those available to date. This is one sense in which the use of corpora is in its infancy. The type of study to be conducted in this book has been, until very recently, virtually impossible because the words we want to look at are not frequent enough to be observed reliably using the relatively small historical corpora that have been available for study.

The relative sparsity of data discussed so far leads to a second point – as the size of the corpora are relatively small, so are the number of points in time when we can observe the words and begin to look at, for example, diachronic change (change through time). Why is this important? If we want to look at something like representation, this may change over time. Also, at any one point in time it may be the case that what we are interested in is simply not mentioned. Again, this distinguishes this type of research from studies focused on grammar – those frequent words and structures relating to grammar occur frequently and reasonably steadily meaning that the study of some features with relatively small datasets is quite possible, as noted. However, if we are interested in a person, a group or an idea, for example, it may or may not be mentioned at certain times, in certain types of texts and by certain authors. Let us look at another historical corpus to illustrate this point.

The Lancaster Newsbooks Corpus is composed of newspapers published in Britain in the 1650s. For the purpose of this experiment we will look at titles published in 1654. This allows us to search just under a million words of data. If we look for the name *Cromwell* we see something of interest. Though he is the foremost politician of his age and the corpus contains 105 mentions of him,

there are seven newsbooks that do not mention him at all. This illustrates nicely the point that, at any given moment in time, even those who are frequently mentioned may not be mentioned by all types of publications or authors. If we look for the name of the battle which effectively ended the English Civil War, we find but one mention – and it is of a ship called Naseby rather than of the battle. We are looking at the wrong point in time and/or the wrong type of text to find mentions of the battle. Newsbooks were reporting news – by 1654 Naseby was old news and hence was not mentioned in the newsbooks. It may be that if we had a corpus of military histories of the time, we may well find a mention of the battle, but we do not find it in newspapers. This neatly illustrates the second point – mention can vary over time. What may be frequently mentioned in one type of text at one point in time may not be mentioned frequently or at all at another, even close, point of time.

The existing historical corpora that have been discussed fall foul of problems such as those outlined when we try to use them for looking at representation through time, for example. As noted in Table 1.1, if we want to look at the word *prostitute* through time using some of these corpora there are very few examples – if any – to work from. Where corpora are better organized to allow specific points in time to be explored in sequence, the gaps between the points in time sampled is so great that a very partial picture of the development of a word may be gleaned. Consider the Brown family of corpora. These are one million word samples of British or American English taken at specific points in time. So far, four points in time have been sampled – the early 1930s, the early 1960s, the early 1990s and 2006. At each sample point a corpus of British English and a corpus of American English have been constructed. Each corpus is approximately one million words in size. Each corpus samples a number of texts of different types written by a variety of authors. They are very useful for looking at slow-moving change in language as discussed earlier in this chapter. Such corpora produce snapshots of the language which, when run together, might be likened to slow motion cinematography – allowing a very slow change to become visible. Yet when such corpora are used to look at non-grammatical features of language change, while they are useful, they suffer from some of the limitations discussed so far. They are relatively small, so only frequent features may be explored. They also sample very few points in time relative to the time period over which language is changing. So, with this family of corpora, we have four sample points per language covering a span of nearly eighty years. As a consequence, it is possible that, in the gaps between the sampling points, change is happening that we simply cannot measure because we have not sampled in

that area. Millar (2009) and Leech (2011) have debated such issues with the Brown family of corpora. While their use is clearly appropriate for looking at grammatical change and the change in frequent non-grammatical features (e.g. Baker 2011), there is little doubt that much is missed that may be of interest to the linguist and historian where we represent such a wide span of time with so few sample points. Arguably worse still, there is a great temptation to draw graphs on the basis of such corpora that imply steady changes between the points when, in fact, we have no real evidence to claim such a steady change. Consider the graph in Figure 1.1.

Figure 1.1 A word explored through the four British English corpora in the Brown family.

In this graph it is implied that the word starts from a low frequency, increases steadily through to 1961, then declines to a plateau between 1991 and 2006. What word is this? It is the word *Hitler*. This graph is almost certainly wrong. It is very likely that the peak of mention is much, much earlier than 1961 and that there were fluctuations in each year. The graph looks artificially neat because of the limited number of sample points and the implied smooth transition of frequency from one point of time to another. We would emphasize again that for gradual change these assumptions may be quite reasonable to make. But for topics or mentions of individuals in a language it is almost certainly the case that relying on so few sample points to cover such a long stretch of time is likely to give highly misleading results. To make the point clearer, consider the graph in Figure 1.2. It is taken from the Times Online, which allows access to all copies of the *Times* going back to the late eighteenth century. It is a graph of mentions of *Hitler*, by decade, in the *Times* between 1930 and 2008.

Figure 1.2 *Hitler* explored through the Times Online, from 1930 to 2008.

Comparing the graphs in Figures 1.1 and 1.2 directly is obviously difficult – we would be comparing corpora of fixed size and a range of genres to a corpus composed of only newspapers with the volume of print produced at any sample point varying. That notwithstanding, the graph in Figure 1.2 tells us something very important – with more sample points we can see more variation in the curve drawn. However, we are still treating a whole decade as one point – what happens if we zoom in further and explore one decade in detail? The picture changes – the number of examples in a decade is not smoothly distributed. Figure 1.3 shows the distribution of the word *Hitler* by year through the Times Online in the 1930s.

Figure 1.3 *Hitler* explored through the Times Online, in each year of the 1930s.

In short what we observe changes as the resolution of our observation improves. The decade-by-decade analysis of Figure 1.2 has value – we discover that mentions of *Hitler* were at their most intense in the 1930s – this accords well with what we know of the history of the twentieth century and as such provides a more credible picture than that in Figure 1.1. Figure 1.2, being based on data from throughout the period, represents the period better than the data in Figure 1.1. Consequently, even though the data is aggregated into decade-long chunks, the overall picture presented looks plausible. Figure 1.3 shows what we get from a higher degree of resolution. Disaggregating the decade into yearly chunks yields dividends – we see that mentions of *Hitler* are not smoothly distributed. From 1933, when he becomes chancellor of Germany, we see mentions of him rise to a fairly high number of approximately 1000 mentions per year from 1933 to 1937. With the Munich Agreement and Danzig crisis in 1938 and 1939 respectively, we see mentions of him escalate as his actions become increasingly newsworthy. All of this accords well with what we might expect, given the history of the period. Yet Figure 1.3 has one last possible contribution to make to this discussion – it is of note that the lowest levels of mention of *Hitler* were in the period in which the Brown family British English corpus for the 1930s was sampled, the beginning of the decade. This probably explains the curious shape of the graph in Figure 1.1 and shows why samples of a small number of points over a long period of time are unlikely to prove fruitful for research questions relating to relatively transient events and mentions of people.

To sum up, another area in which historical corpora must develop is in terms of coverage if they are to be of use for a broader range of research questions. Corpora which cover long periods of time with little data and few sample points are of limited use – or no use at all – for exploring changes in language that may be relatively transitory and dynamic. Trying to explore such features using such corpora can lead to highly misleading results. The question that remains, however, is: what degree of resolution might be necessary to explore a research question? This is bound to be a balance between what is possible and what is desirable. This is a point we shall return to later.

So, in order to use corpora to explore historical questions it seems that a continuous run of material might be desirable. Similarly, it seems that, while we may trade off resolution for sufficient material to analyse, we might want to work with reasonably short spans of time in order to achieve some tolerable level of resolution in the analysis. While the ideal may be to analyse by day, month or year, from the graph in Figure 1.2 it certainly seems that analysing by decade gives useful results. As long as there is not sufficient data to allow day/month/

year resolutions in searches, a heuristic adopted in this book is that a decade-by-decade analysis represents a minimum level of resolution for the purpose of historical investigation. So any corpus we might want to use for our study should at least allow this level of resolution, with the corpus being a collection of material drawn from across the period in question, rather than from discrete and disconnected sample points.

Do such corpora exist? Have researchers worked on them? It is possible to say a very tentative ‘yes’ at this point. If we view a corpus simply as a large collection of textual material that can be searched by computer, then the answer is ‘yes’ – though, as will be shown, what is lost by accepting such a definition is substantial. If, on the other hand, we want that material to be subject to manipulation by various corpus processing tools and subject to close analysis where appropriate, then the answer is more ‘not very often, if at all’. However, this is changing as Section 1.5 and the rest of this book will show. Yet before we explore what is now possible, let us focus first on the weaker definition of corpus linguistics and explore the usefulness of Google Ngrams and Culturomics to the study of language and history. Readers not interested in this may skip directly to Section 1.5 at this point.

1.3 Google Ngrams and Culturomics

Google, the search engine provider, has been scanning books for some time and started to make them available online in 2004. Over time they have amassed what is, in effect, a substantial corpus. They have produced the corpus using optical character recognition – effectively making printed texts into machine-readable ones suitable for rapid searching. This is the data that underpins the Google Ngram Viewer service. When searched for in that system, a word appears in a simple graph, not dissimilar to those in Figures 1.1–1.3. The resolution of the search is to a year and an indication of the frequency of the term relative to the total words in the corpus for that year is given – that allows a comparison of relative frequency similar to that which we achieved by normalizing frequency to occurrences per 100,000 words in Table 1.1. The corpus you search can be looked at using so-called wild cards – characters which can be used to represent single characters or sequences of characters in a word. So, to search for all words beginning with *prost* one would type in *prost** and words such as *prostitute*, *prostitutes* and *prostitution* would be included in the result returned. However, this would also include results for words such as *prostate* and *prostrate*.

Accordingly, the search engine allows you to look for all inflected forms of a word, solving this problem. What if one wanted only to find *prostitute* as a noun? Again, the system can help as you can search for *prostitute* as a noun separately from *prostitute* as a verb. Helpfully, for a book like this, the so-called Google One Million collection of texts contains books from 1500 to 2008 and so represents a potentially useful source of data for historians. However, there are issues with the Google corpora – first, they can only be accessed via the Google interface. The operations you can carry out on the corpus are limited by what is, relative to the tools that will be introduced later in this chapter, primitive. Simply looking up words and combinations of words is of limited interest and utility. First, you need to know the words and word combinations to look for – the capacity for users to discover interesting patterns is limited. Unless you look for a pattern, you may not see it. Yet unless you know the pattern you will not look for it. This is an issue because, as will be shown in this book, as we move into the past our intuitions about language experience strains and breaks. We need to use tools which allow us to discover patterns we may not be aware of – and some of the tools of corpus linguistics, notably collocations and keywords, which are used later in this book, allow for this. We define both collocates (words which co-occur frequently) and keywords (words which are unusually frequent in a corpus) later in the chapter, but for now we need simply to acknowledge their usefulness to corpus linguistics and, by contrast, how unhelpful it is not to be able to explore them. Hence the interface through which the data must be accessed is a major limiting factor.

Secondly, how accurate are the scans of the original texts? We have no way of knowing this as the original texts cannot be downloaded and used. However, it is not unreasonable to assume that there are errors in the scanning of the texts. Also, given the effects of time, the declining quality of print, and the prevalence of variant spelling in the early modern period, all of which must have impacted upon the quality of the OCR process, it is not unreasonable to assume that, in the absence of evidence to the contrary, errors are not smoothly distributed in the Google One Million corpus and that they are skewed towards the earlier data.⁸ Research on other large-scale OCRed resources suggests that the accuracy of such scanning may be low and get lower the further back in time you go. Tanner et al. (2009) undertook a study looking at eighteenth- and nineteenth-century material OCR scanned by the British Library. On average, 78 per cent of the words in the nineteenth-century material were accurately identified, while 65 per cent of the words were, on average, correctly identified in the eighteenth-century material. It would be surprising if there was not to be a broadly similar scale of error in the Google material of a similar age.

Thirdly, the lack of access to the original texts is a major drawback. Without being able to see the original examples in context, or even knowing which texts are used in the corpus, it is difficult to reliably interpret the meaning of any word in context. The limited contexts made available by Google in the form of ngrams is not adequate for this purpose. While concordancing, collocation and keywords can be helpful and also often present limited views of the data, the ability in corpus analysis tools to navigate easily and swiftly to the full text underlying an analysis is critical in terms of aiding analysis. At best, Google Ngrams allows one to navigate to view a list of the books in a sub-period of the analysis. From there you can scroll down a list of books in which the word occurs in that period, with results presented in much the same fashion as if one had searched for something via the Google search engine. This is an approach to the presentation of the data that is clumsy and not fit for the purpose that a corpus linguist or historian would put the corpus to.

We will not labour these points, but they are important. Perhaps the best way of demonstrating this is to look at culturomics, a research area which has enthusiastically embraced such data in order to make sweeping statements about language and culture through time. Spurred by a milestone paper by Michel et al. (2011), culturomics has become something of an industry with researchers, typically not linguists, but rather from the physical and other sciences, taking the ngrams available from Google and using statistical techniques that they are familiar with in order to make statements about culture through language. While we are sympathetic to the goal of studying culture through the written record, the way in which culturomics does this sets it aside sharply from corpus linguistics. As noted, corpus linguists would typically manipulate textual data using a range of tools which have a proven ability to give us insight into language and through language, where we wish to explore it, culture. These tools are not available to the culturomics experts, for the reasons outlined. Similarly, the study of the results and the working back to the texts themselves, a well-established procedure in corpus-based discourse analysis, in particular (see, e.g., Baker et al. 2008; Partington 2003), is difficult or well-nigh impossible to operationalize using the Google Ngram Viewer. It is perhaps for that reason that culturomics researchers do not seem to make any notable effort to link their analyses back to close reading of texts. Another reason they may not do this is that they do not value this as a process. Given that they are typically not linguists or, where they are, they have not worked with linguistic corpora previously, they make a series of leaps that corpus linguists would find surprising. First, as noted, they do not link their analyses back to close readings of texts. Secondly, they do not use the

procedures which corpus linguists have established as fruitful in the exploration of culture via language. But thirdly, and perhaps most damagingly, they do not seem to conceive of the mutable nature of language, especially when studying representation and meaning. In this important sense language varies very significantly from the physical world to which many culturomics researchers have devoted their studies. It is perfectly reasonable to assume that, if you could study hydrogen atoms through the centuries, they would be no different between the Big Bang and the present. Similarly, a hydrogen atom a million light-years away from Earth would have the same properties as the hydrogen atoms on Earth. Yet such statements are simply not true for language. Words vary – they vary through time (diachronically) and they vary across space at any moment in time (synchronically). Words inhabit a social world. Different people and groups may use the same token (word) in the same way. To give an extreme example, the word *nigger* is unacceptable when used by some speakers in some contexts in present-day English, yet when used between certain types of speakers in other contexts it is acceptable. Its meaning changes by context and speaker. Similarly, words change through time. What is true of the word *nigger* today was not true of it in the past, when, no matter how distasteful it may seem to the modern reader, the word was not thought to be as unacceptable as it is considered today. So, in drawing graphs that span centuries, one is indeed looking at the life of one token, one symbol to which meaning is attached. But unless we want to limit our claims to simply saying ‘the combination of letters *prostitute* has the following frequency through time as shown by this graph’ then we have a problem. What we are measuring, while it appears to be immutable, is almost certainly changing through time – meaning is protean. Without access to appropriate tools to study this change of meaning, we may be proceeding on the almost certainly false assumption that the meaning of a word is as immutable as the composition of an atom. It is not. We are also left in a position where, while we may marvel at the peaks and troughs of a frequency profile for a word or phrase over centuries, we are not really any closer to describing and explaining the phenomena that underlie that profile. Culturomics is theory and narrative free – and at best the data it produces is a challenge to theories and historic narratives, not a replacement for them. Let us illustrate these problems by looking at a study in culturomics.

Let us consider the work of Gao et al. (2012), a group of engineers and mathematicians who have published in the area of culturomics. In Gao et al. (2012), using a lot of mathematics that we need not concern ourselves with here, they look, via Google Ngrams, at words which describe social and natural

phenomena in the nineteenth and twentieth centuries. For example, social phenomena include *communism*, *Christian*, *democracy*, *famine*, *politics* and *war*. Natural phenomena include *avalanche*, *earthquake*, *fire*, *tornado* and *typhoon*.

The paper founders almost immediately on a linguistic point – not every use of each of these words relates to the meanings that the authors believe that they do. Take *avalanche*, for example. Using a corpus of 1.1 billion words of present-day English⁹ we looked at the word *avalanche*. Of the first fifty examples of the word, an exploration of the context in which the word occurs demonstrates that the use of the word *avalanche* to refer to the natural phenomenon applies to only twenty-eight of the fifty examples. The other sense of *avalanche* is metaphorical, relating to being swamped by a sudden flood of something, at least in this brief study. There is no discussion of such issues in the paper, even though it touches on almost every word they look at – other words that they study may also be used as metaphors and similes but there is no discussion of this. Some of those words are also bound to be sensitive to historical events; it is reasonable to assume that around the time of the world wars the literal use of *war* would have increased. Such issues may have been shown had corpus linguistic techniques been used. In another example, the word *political* occurs very frequently with the word *earthquake*, a clear sign that the word has a metaphorical use which is clearly social, yet the authors would have only considered it with a literal meaning which they would have considered to be physical. Given that the paper is essentially using word frequency while assuming that each word has a meaning that consistently groups it into either the ‘social’ or ‘natural’ phenomena groupings, it is not unfair to say that the results are flawed because the authors misconstrue what they are counting – they are not counting words coupled with immutable meanings. They are counting words which associate themselves with different meanings over time and in different contexts. In a newspaper article on an election, the meaning of *landslide* is not the same as it is in an article describing a natural disaster. In a period during which a war is raging, the meaning of *war* may indeed primarily refer to armed conflict. In a period of peace that is unlikely to be the case. The paper states that ‘social and natural phenomena are governed by fundamentally different processes’ (Gao et al. 2012: 1956). This would hardly come as a surprise to any social scientist or humanities researcher. Yet it is certainly something to bear in mind when scholars versed in studying natural phenomena turn their hand to studying social phenomena.

While interesting, the linguistic and historical naivety of much culturomics research hugely reduces its value. Note that we are not saying it has no value – there is value to be gleaned from looking at word frequency data, as this book

will hopefully show. However, a result based on word frequency data alone, especially results premised on simplistic views of word meaning, is likely to be distracting at best and dangerous at worst.

However, it is still the case that the data underlying the culturomics venture could be useful for historians if they could access and manipulate it easily – though for the purposes of this book it may be less useful. As discussed earlier, there is cause to believe that the further back in time one goes with the Google material, the greater the noise caused in the analysis by OCR error. The www.culturomics.org website certainly seems to caution that as one moves back in time with the corpus the results are less reliable, though the reason is ascribed to data sparsity caused by the sampling of the books in the earlier centuries when fewer books were printed: ‘Watch out for the time period your [*sic*] are looking into: the best data is the data for English between 1800 and 2000. Before 1800, there aren’t enough books to reliably quantify many of the queries that first come to mind.’¹⁰ Whatever the cause of the issues with the Google material for earlier centuries, there is little doubt that the problem exists.¹¹ Consider the graph shown in Figure 1.4 for the word *prostitute*:

Figure 1.4 The word *prostitute*, 1600–2000 in Google Ngram Viewer.¹⁷

It seems highly implausible that the word *prostitute* was proportionately at its most frequent in English in the 1660s, as the graph in Figure 1.4 suggests. This is almost certainly an example of skew in the corpus – in one of the years of the 1660s it would seem that there is a text, or texts, which use the word *prostitute* unusually frequently. Given that, as culturomics.org warns, there are fewer texts per year in this part of the corpus, the opportunity for a small number of texts to produce skewed results like this is enhanced. It is difficult to discern exactly

how this spike occurs using the rather basic interface Google provides to look at the examples making up the count of usages. Hence, this example shows both the dangers of using the Google material to explore the seventeenth century and how important it is to be able to easily navigate back to the examples producing a feature in a graph so that they may be fully understood. For now we will leave our discussion of the Google data, though we will return to it when we look at the word *prostitute* in the seventeenth century using other corpus data later in this chapter.¹²

1.4 An important corpus – the Old Bailey Corpus

So far the focus of this book has been on issues in the use of corpora for historical research. Existing corpora will not permit many of the studies we may wish to pursue. Some corpora, notably the Google data, seem to provide a solution but are bedevilled by a restrictive interface and issues with data prior to 1800.

Another corpus which might hold some promise for the purposes of this book is the Old Bailey Corpus.¹³ This is a corpus derived from the texts produced for the Old Bailey Online website – a web-based concordancer, of sorts, that allows one to search through the proceedings of the criminal court, the Old Bailey. In principle this is very useful data, but is less useful for this book – our aim is to look at the representation of prostitution in general, not simply the representation of prostitution in criminal proceedings during which, it is reasonable to assume, it will be presented as wholly problematic. Another problem with the Old Bailey Corpus is its coverage – it covers 1674 to 1913, whereas we wish to study 1600–99. A final problem is the scale of the corpus – while it contains 2163 volumes of proceedings resulting in a corpus of almost 134 million words, the corpus addresses one objection raised earlier in the chapter, yet falls foul of another. It is a continuous sample through the period covered by the data – one can, in principle, achieve resolution at a number of levels as the court sat fairly continuously in the period covered by the corpus. However, relative to the span of time covered, the size of the corpus is not overwhelming which probably, once again, means that it is best for the study of frequent features – grammatical features or, perhaps, concepts frequently addressed in the courtroom. So while the corpus is without doubt very useful, it is not the ideal corpus to use for this book. Setting aside the issue just discussed, the focus of this book is on the seventeenth century, but the Old Bailey Corpus

only covers the latter quarter of it. For these reasons, this corpus is not central to the study undertaken in this book.

The Old Bailey Corpus is but one example, of course. There are other historical corpora which can be used, perfectly satisfactorily, to study specific research questions. If these questions rely on specific genres, it is not unusual to find small, focused, corpora which allow an investigation of that genre. For example, the Lancaster Newsbooks Corpus (Hardie and McEnery 2009) is useful for those interested in journalism in the 1650s and the Helsinki Corpus of Early English Medical Writing¹⁴ is valuable for those looking at the development of both medicine and the genre of medical writing. However, as with the Old Bailey Corpus, such corpora are of use to those interested in a narrow research question. The challenge we set ourselves for this book was to find a very large, general corpus of written English for the seventeenth century. Hence, while the focused historical corpora are very useful, they do not suit our purpose generally, though we will use the Old Bailey Corpus in Chapter 3 of this book to explore some criminal cases against prostitutes. But we cannot use it to look at public discourse; for that, we need to look at a wholly new corpus.

1.5 Introducing the corpus to be used in this book

At this point we think it is likely that any reader of this book will have been intrigued by the possibility of a corpus-based approach to historical research questions but may be beginning to despair that it is possible. For the purposes of historical investigation as we have outlined them, the resources discussed so far are deficient in a number of respects. At this point, we could argue for a change of scope of our study (e.g., looking at prostitution cases in the Old Bailey Corpus), shift forward to the modern day and demonstrate with modern resources what might be possible if we had access to similar resources for previous centuries or, perhaps, simply decide instead to write a paper critiquing the state of the art in historical corpus linguistics. Fortunately, we need do none of those things. A new corpus is being constructed which meets all of the criteria we have outlined for our corpus-based investigation of historical questions. It covers the time period in question. It is large – with approximately one billion words of data available for the seventeenth century. It samples from all of the years of the seventeenth century. Finally, it is composed of printed materials of all sorts, not simply books. The corpus in question is the Early English Books Online (EEBO) corpus.

1.5.1 Early English Books Online

Pioneering work scanning and putting online digital facsimiles of works from the past has a history that predates Google Ngrams and an important part of that history was the creation of EEBO by a company called ProQuest. These books proved an immensely valuable research resource, yet it was of limited interest for linguists as the texts were images rather than machine-readable texts. In 1999, however, the Universities of Michigan and Oxford, in collaboration with ProQuest and the Council of Library and Information Resources, began the text creation partnership with the express goal of producing fully machine-readable texts from the EEBO collection. The text creation partnership grew and now over 150 universities worldwide are contributing to this effort. The texts are quality-controlled transcriptions of original texts. At the time of this book going to press, 40,000 texts are available in the EEBO corpus.¹⁵

The EEBO, as a corpus, has some distinct advantages over the corpora behind Google Ngrams. First, we know which texts are in the corpus¹⁶ – this is not true for the Google material. This gives us a better insight into possible issues of skew in the corpus. Secondly, we know the quality control processes used to create the data and can check that as we can gain direct access to the texts in the corpus. Thirdly, we can use the corpus and manipulate the data in powerful corpus searching engines which allow all of the tools and techniques developed in corpus linguistics to be applied to the data. Fourthly, because the corpus can be manipulated directly, the corpus can be processed. As a result, at Lancaster University we have been able to both part-of-speech tag the corpus and use powerful software to deal with the issue of spelling variation in the data (see Baron et al. 2009). Fifthly, as with the transcription, because we have access to the corpus itself, it is possible for us to explore results in order to see if any of them are skewed unduly by errors in the part-of-speech tagging or variant spelling handling. Finally, we get scale – the EEBO corpus material for the seventeenth century amounts to nearly a billion words. As this is fairly smoothly spread through the century, we can hope to gain resolution to at least the decade level.

In short, the EEBO corpus is close to ideal for the purposes of this project. It gives us direct access to the texts and, because the texts can be manipulated through concordancing software, an easy ability to map between results and the texts that produced those results. This in itself enables the type of supporting close reading for results that both historians and corpus linguists would expect. It is possible that readers may think that some of the advantages outlined above

are mere adornments – potential advantages that rarely make a difference. We can say with confidence that this is not true and that, as the results in this book unfold, the value of these advantages, and the corresponding disadvantages of the Google material, will, we believe, become apparent.

One immediate point can be made based on the scale of the EEBO corpus – there is sufficient evidence to allow the type of investigation we want to undertake. To illustrate this, consider the data shown in Table 1.1. Three historical corpora yielded two examples of the word *prostitute*. In the EEBO corpus, we have 2,397 examples of the word in 1,616 texts. Throughout this book we will be using the CQPWeb system as provided to us by Dr Andrew Hardie at the University of Lancaster. CQPWeb is a powerful online concordancing system which allows us to use state-of-the-art processing tools to swiftly and reliably manipulate the billion words we are interested in. Rather than simply drawing crude graphs, as the culturomics website permits, CQPWeb provides useful information in a variety of forms. For example, Table 1.2 shows the useful distributional data provided for a search such as that made for *prostitute*.

The data in Table 1.2 compares very favourably indeed with the graph of the word reproduced in Figure 1.4. While visually appealing, the graph is nothing more than a visual rendering of a normalized frequency. We have no idea of how widely spread mentions of the word are, or how many words are at each sample point in the graph. In Table 1.2 we have a much richer context in which to begin

Table 1.2 The distribution of mentions of *prostitute* in the EEBO corpus in the decades of the seventeenth century

Period	Words in the corpus for this period	Number of mentions of <i>prostitute</i> in this period	The number of files for this period in the corpus that mention <i>prostitute</i> at least once	Normalized frequency per million words in the period
1600–9	57,272,438	95	69 out of 1349	1.66
1610–19	61,922,837	146	92 out of 1465	2.36
1620–9	55,748,866	118	88 out of 1674	2.12
1630–9	63,496,401	166	121 out of 1572	2.61
1640–9	87,480,996	196	147 out of 7424	2.24
1650–9	168,912,439	338	216 out of 4690	2
1660–9	111,998,646	279	179 out of 4227	2.49
1670–9	118,167,747	265	175 out of 3920	2.24
1680–9	142,071,417	408	282 out of 6916	2.87
1690–9	128,494,904	385	246 out of 5588	3

to interpret and account for the word *prostitute* in the seventeenth century. This is just the beginning – because CQPWeb is capable of so much more than producing the data that can be seen in Table 1.2, impressive as that is.

Before leaving the discussion of EEBO, it is probably worth noting that this corpus is probably more unfamiliar, at the time of writing, to linguists than it is to scholars in the so-called digital humanities. In the digital humanities there is a growing body of research on EEBO and other historical text collections. Some of this focuses on what is lost and gained by shifting data from the physical to the digital (see, e.g. Kichuk 2007). Withington (2013) has used very basic corpus-based methods in order to introduce an analysis of the semantic complexities of the term *peace* in early modern England. He has used the Shakespearean¹⁸ and EEBO corpora in order to compare the number of times *peace* appears in each corpus and, specifically the title pages of EEBO, in comparison to other search terms, *war* and *god*. This is a rudimentary use of corpus linguistics and mainly serves as an introduction to Withington's subsequent sophisticated qualitative analysis but it is, nevertheless, an example of how historians are starting to make use of digitized sources and, moreover, are becoming aware of how corpus linguistics can benefit their research.

A growing area of focus is upon using techniques from computer science to try to automate various aspects of the analysis of historical materials, not entirely analogous to the culturomics approach to language outlined, but far from different. This work is not based upon linguistic insights into language. Nor did it find its genesis in a wish to meet the need of historians studying large historical text collections. Rather, techniques such as topic modelling (e.g. Wang et al. 2012) have essentially been repurposed from work in information retrieval and library science, designed to service the needs of librarians in the 1970s, 1980s and 1990s. Work on named entity recognition (e.g. Grover et al. 2008) grew out of work in the 1990s, often funded by the American military research agency, DARPA, in the United States, which had perhaps understandable interests in looking at patterns of mention of named entities relating to such areas as finance and defence in newspapers and other texts.¹⁹ Using such approaches researchers have used automated approaches to study subjects such as conceptual history (see, e.g. Pumfrey, Rayson and Mariani 2012), perceptions of the Second World War across different text collections (Odijk et al. 2015) and textbooks aimed at children in German schools in the period 1850 to 1918 (Schnober et al. 2015). This book takes a different approach. Our aim is not focused on automation. Our goal is to bring linguistic insights to the study of language and discourse in the seventeenth century. The

tools we use to do this are discussed in the following section. But for now it is important to note that our focus is upon bringing corpus linguistic insights and historical insights together. The result is machine-aided, rather than purely computational, analysis with a sound footing in concepts such as collocation which are of proven utility and theoretical value in linguistics. While a critical overview of the approaches rooted in computer science is not a goal of this book, we do, in Section 5.3, evaluate a tool which may have allowed an automated content analysis of *EEBO* to be undertaken. It is perhaps the case that the critical investigation there may suggest the need for a more critical engagement with automated analytical procedures in general.²⁰ However, as noted, that is not the goal or aim of this book, so let us return now to consider how we will exploit the *EEBO* corpus.

1.6 Tools and methods – How we dig into data with CQPWeb

So far we have alluded to some of the processes of which CQPWeb is capable. As we move towards the conclusion of this chapter, now is the time to review the tools that CQPWeb provides that will be used in this book. Readers familiar with corpus linguistics may safely skip this section.

Collocation is a central method for the exploration of word meaning in corpus linguistics. The idea of collocation is old (e.g. see McEnery and Hardie, 2011, for a fuller discussion of collocation), but began to form into a shape in which it had a clearer operational definition following the work of John Sinclair in 1970 (republished as Krishnamurthy et al. 2004). The insight that Sinclair gave is that certain features seemed to be important in terms of how words co-occur. Words co-occurring with one another in quite a narrow span, for example, five words either side of a word we might be interested in investigating, seem to be strongly linked to the decoding – or potential creation – of meaning. Measures have been developed for exploring words co-occurring in such a window in order to allow us to systematize the selection of words which seem to keep company with a word more frequently than we may expect – so-called collocates. To identify a word as a collocate, we can look at the frequency of co-occurrence of the collocate with the word we are interested in and contrast that with the frequency with which the word occurs with other words. We can subject those frequencies to a statistical test to ascertain whether the word has a statistically significant attraction to the word we are interested in or not (this provides, in

essence, a simple yes or no result, it is either significant or not). We may also check those frequencies to see how strong the difference is. So as well as being statistically significant, we may find that the so-called effect size – how marked the difference is – will vary between the statistically significant collocations. In this book we will use a statistic called the Log Ratio statistic²¹ which, helpfully, combines a test of statistical significance with a measure of effect size – so, for the words above the significance threshold for the measure, words which get a higher Log Ratio score exhibit a greater effect size than for those words with a lower Log Ratio score.

This may sound rather abstract, so let us consider a simple example. The word *bank* has multiple meanings. However, in context, collocations make the meaning associated with the word in that context manifest to the hearer or reader – in the British English 2006 corpus (Baker 2009), the word *manager* is a collocate of *bank*. In an example like ‘the salami sandwich that the bank manager had been planning to have for lunch’²² the word *manager* helps the reader to identify the word *bank* as meaning a financial institution as the collocate *manager* helps select that meaning of *bank* – river banks do not usually have managers. Used in another way, collocation can help to discover word meaning; hence, it has been a key tool for lexicographers using corpora. Pretend that you did not know the word *bank*. If we look at two million words of present-day English²³ we find collocates for *bank* that show us, when we explore some examples of the words used in context, that banks deal with money (e.g. *money*, *stock*) and the banks that do so have names (e.g. *european*, *midland*, *scotland*) but we also find that they are locations by rivers (e.g. *west*). Collocates can help us to discover nuances of meaning of words that we think that we know, or they may allow us to explore meanings of words which we do not know at all. We mentioned here the need to explore context to understand collocates – corpus linguists do that through concordancing, a process whereby examples of a word are shown organized as one example per line, with the word of interest in the centre of the screen. This arrangement greatly facilitates the process of looking for patterns in the data by a focused closer reading of the text.

An important feature of collocates is that they may show semantic preference and discourse prosody. Semantic preference occurs when a word collocates with words of similar meaning. Discourse prosody occurs where a word attracts other words which are either positive or negative in meaning. In this book we will largely bypass the term ‘semantic preference’, though the concept does underlie the organization of collocates into semantic fields which will be a key feature of the analyses undertaken in Chapters 5 and 6. Discourse prosody is

also important in that chapter, as it will be shown that some of the words we investigate were routinely associated with negative concepts/ideas. Discourse prosody is arguably the most subtle way in which collocation encodes meaning in that it is indirect, the cumulative effect of the pattern of collocates that a word attracts. Consider the word *cause*. The corpus linguist writing this book has asked many students in seminars and lectures over the years to write a short definition of this based on intuition and the answers produced rarely vary. Typically, a definition something like ‘to make something happen’ is provided. However, as Stubbs (1995) has shown, in fact what is caused to happen is negative in the overwhelming majority of cases in spoken and written present-day English. So, in the British English 2006 corpus the strongest collocates of *cause* are *damage* and *death*. While we could, in principle, say ‘cause happiness’ our habit is to not do so because, corpus linguists would argue, the negative discourse prosody of *cause* is part of the meaning of the word.

One final concept to explore under the broad heading of collocation is colligation. Collocates are those words which help to define meaning through co-occurrence – colligates, on the other hand, key more directly into grammar. Co-occurrence with grammatical words such as ‘articles’, ‘prepositions’ and ‘subordinators’ is the most obvious manifestation of colligation. In the studies in this book we will typically focus on collocates and will largely set aside colligates unless the discussion of them is relevant to the study of prostitution.

Another technique that will be used in this book is the keyword procedure. Let us explain simply what the keyword approach does. Imagine we have two corpora, A and B. The keyword procedure checks, for each word in corpus A, how frequently that word appears in corpus A relative to corpus B. As with collocation, the keyword may be marked by statistical significance and effect size – hence once again we will use the Log Ratio statistic in this book. Note that in calculating the test of whether or not a word is key, the statistics takes corpus size into account: if corpus A is twice the size of corpus B, then if a word is twice as frequent in A as it is in B, this should be the cause of no excitement – the relative frequency of the word in both corpora is the same. Finally, note that keywords may be positive (the word has increased in frequency) or negative (the word has decreased in frequency).

Keywords are an important technique in corpus linguistics as they help us to grasp the ‘aboutness’ of a large text collection. If we find a positive keyword, a word is being mentioned more often in the corpus in question, relative to the one being compared to, in a way that is unlikely to be explained by chance. Similarly, a negative keyword occurs when a word is being used less in a corpus, relative

to the corpus it is being compared to, in a way that is unlikely to be explained by chance. Keywords have proven their value in looking at large corpora which span a long period of time, notably in the study of Gabrielatos et al. (2012), where keywords allowed different British newspapers' attitudes to Muslims to be explored in a corpus of 140 million words of newspaper articles mentioning Muslims or Islam in the period of 1998–2009.

The challenge of looking over stretches of time is only amplified in this book by our decision to look at a century of material. These challenges will be greater, at times, than those faced by other studies. For example, Gabrielatos et al. (2012) worked with a corpus in which they could reasonably produce a resolution of analysis as small as one day in their study of newspapers in an eleven-year period. This led them to consider methods to look for trends in that data as, with many more sampling points than other studies mentioned in this chapter so far, they had a much greater issue when it came to identifying trends in the data and identifying points that varied significantly from that trend. We too will face the challenge of looking for trends in a long period of time. A greater challenge will be set by collocation. Gabrielatos and Baker (2008) noted that collocation was subject to variation in discourse over time. They developed the concept of consistent collocates to deal with this, focusing on collocates which appeared reasonably consistently with a word across the points sampled in their study. Studies like these will provide a useful guide to this book as we deal with similar issues on a larger scale. Yet the scale of our enterprise, across a century, will present other issues to which a methodological response is required. In the sub-section that follows, we will briefly outline a series of methods we either use or have developed for this book which revolve around the use of collocation to explore four states – consistency, initiation, termination and transience. Note that we may apply these labels either to individual collocates, or to the semantic fields which the collocates are drawn from (as happens in Chapter 5). In the discussion that follows, however, we will focus on individual collocates to illustrate the idea. We will return to these concepts later in the book, after they have been used in Chapter 5, to try to understand how we may explain such patterns of collocation.

1.6.1 Consistency

In this book, we will be looking principally at collocation and the extent to which it intersects with prostitution. Collocations will form the evidence both for any claimed shifts in word meaning that we discuss and for any changes

in representation. The last point is an important one. When we see collocates change in the life of a word, what are we seeing? There are a number of possible answers and only by looking at a collocate through time can the most plausible answer be determined. Consider a word, let us call it A. Over 100 years, in an analysis with a resolution of a decade, we see A collocate with B in every decade, but with C in alternating decades. It would be unusual to claim that this was evidence of the semantics of A changing. Rather, what we are observing, would more probably be a word A which has two collocates, B and C. B is a collocate which does not seem to be prone to fluctuation, while some factor seems to be making C enter and leave discourse. Yet throughout we would claim that A has meaning formed in collocation with B and C – not that the meaning of A shifts every ten years to either include or exclude C. Consider the word *landslide* as discussed earlier in the chapter. It has collocates based on its metaphorical uses and collocates based on its literal meaning. Landslides in the UK are rare and hence, unless a major landslide occurs and is commented on, the word typically appears as a metaphor. However, when a landslide is reported, it is not that the word gains a new set of collocates, and with it meaning. It is, instead, that a latent set of collocates are selected that are appropriate to the discourse. Hence when looking at consistent collocates, Gabrielatos and Baker (2008), rightly in our view, came to the conclusion that to be consistent a collocate need not always occur with a word. There may be periods during which discourse makes a set of collocates latent, or at least infrequent for a word. Collocates are sensitive to the discourse environment which, in turn, is sensitive to the social, cultural and moral pressures being exerted on language at any one moment in time.

Consequently, consistent collocates in this book will be those that occur in at least seven decades out of the ten being looked at (the rationale for this is discussed in Section 1.6.5).

1.6.2 Initiation

A clearer sign that a change worth investigating has occurred in language is marked by the initiation of collocation. If a collocate has not been observed in a diachronic corpus until part way through it and then appears consistently for at least three consecutive decades, then we deem this collocate to be a candidate initiation collocate. In short, there is sufficient evidence that this may be a collocate coming into being and that a close examination of the context and prior context of that word is warranted to confirm this. So, imagine that in the

period 1600–40, word A collocates with B and C. But in the period 1640–99 it still collocates with B and C but, from 1660 onwards, it also collocates with D. D would then be looked at as a possible initiating collocate. The reasoning is, we hypothesize, that several decades would be a long time for a collocates to remain latent in a speech community. At the very least, if this was the case, you would have to explain how a latent collocate was communicated to new members of a speech community. Wrigley and Schofield (1981) calculated average life expectancy in the UK to be around forty years of age in the mid-seventeenth century. Imagine that a speaker is born in 1600 and that they have an average lifespan for the age. By the time that the collocate D appears in the 1640s, it would be difficult to explain how this speaker could have contributed to this collocate's survival if it was not a notable collocate in their lifetime. It seems more likely that this is a collocate coming into existence. Yet, note that we have used the phrase candidate initiating collocate. There may be good reasons to believe that this collocate could have been current but absent from our corpus, so when we observe the collocate for the first time it should be carefully examined. For example, it may be the case that this collocate was frequent in speech but not in writing. This would explain how it had, in fact, been a collocate in the lifetime of our imaginary speaker of Early Modern English – it was part of their spoken repertoire but was absent from writing in their lifetime. By exploring the context in which the collocate comes into existence, we may see evidence of that; for example, it may appear initially in quotation. So while the collocates may be suggestive of a collocate entering the language, we should approach the category of initiating collocates with caution and curiosity, taking nothing for granted.

1.6.3 Termination

Terminating collocates are very much the inverse of initiation collocates. Where we have seen a collocation that has been consistent in a corpus disappear and not reappear over a sustained period of some decades, we have a candidate terminating collocate. In this case it is suggestive that the collocate is entering either a long period of latency or it is entering a period of latency as a prelude to being abandoned. Again, however, the context in which this occurs should be explored. As with the initiation discussion, it may be that the collocate is simply shifting to become a feature of speech rather than writing, or is now becoming closely associated with a genre not represented in the corpus. However, equally, as with all collocations, there may be historical explanations which explain well

why the collocate is either coming into or going out of use. Those historical explanations seem to us to be the key, in particular, to the next category of collocate, transient collocates.

1.6.4 Transience²⁴

Transient collocates we define as collocates which are associated with a word for a short period – in our terms a decade or two – before they vanish. We think it highly likely that these collocates are well worth exploring as being diagnostic of transient shifts in society, culture and morals. As attitudes change, events unfold and the cultural practices of a society shift, we should expect to see language shift with it. Some changes sustain – the collocates are thus initiating collocates. However, some collocates are drawn into usage and either terminate or lapse into latency, though they may re-emerge at a later date. Hence, in a century a transient collocate may appear and disappear on a number of occasions.

1.6.5 Measuring collocates and keywords

How do we identify the collocates outlined above? First, we need a procedure for identifying collocates in the corpus. Then we need criteria for categorizing collocates as consistent, initiating, terminating and transient.

All of the measurements used are, to some extent heuristic – the settings for collocation are normally a window of $\pm$ five words and we will use that in this book also. We use the Log Ratio statistic for the calculation of the score, which determines whether an association between two words is counted as a collocation or not. As noted this is a relatively new statistic which has the merit of testing both significance (i.e. whether it is possible to ascribe statistical significance to the frequency with which two words are attracted to one another) and effect size (i.e. how strong is the bond between the two words).

Consistent collocates have been measured by Gabrielatos and Baker (2008: 11) as being present in seven out of ten cases tested. The resolution used for their study was years, but we shall use the same measure with a decade level resolution.

Initiating collocates will be determined, for the purposes of this book, as collocates which are not present in our corpus from its beginning for at least forty years. From the point it appears it occurs in at least 50 per cent of the remaining decades, with a minimum of two occurrences. Terminal collocates

will be determined to be those collocates which are present in at least 50 per cent of the decades from the start of the corpus, being a collocate in a minimum of two decades. It will exit the list of collocates for a word for a minimum period of forty years. Transient collocates are those which do not meet the criteria for being classified as consistent, initiation or terminal collocates.

These measures are a starting point for the investigation of collocates to determine whether the classification given to them is plausible linguistically and historically. Note that the definition, coupled with the resolution at decade level and the span of 100 years, means that some of the categories, notably termination and initiation, are bound to be distributed unevenly through the decades. Initiating collocates can appear no sooner than the 1640s by our definition. Similarly, terminating collocates can occur no earlier than 1630 in our data. We have set our parameters for these categories conservatively – it may well be that collocates may be initiated or terminated more swiftly than we imagine. However, in both cases we have used the average lifespan as the, admittedly somewhat arbitrary, filter through which termination and initiation must pass, while also applying a weak consistent collocate test of at least 50 per cent (as opposed to the full consistent collocates measure of 70 per cent).

The heuristic nature of the investigation is again underlined by our decision to work with conventional decades – these are arbitrary units. We could well look at ten-year chunks running from 1601 to 1611 and so on rather than starting with 1600 to 1609 as we do in this book. It would be interesting to see whether other approaches to looking at change over time might be supportable – for example, looking at a sliding window and calculating changes in collocations on that basis, starting the analysis at 1600–9 and then slowly changing the window to 1601–10, 1602–11, 1603–12, etc. While this form of analysis is not yet supported by corpus searching software and is extremely laborious to do manually, in order to see whether such an approach may have a dramatic effect on our findings we did run a sliding window experiment with the word *whore* between 1600 and 1620. The results were interesting but not alarming – the collocations, while they change across the period, were reasonably stable when you compared one ten-year window (e.g. 1600–9) to the next (in this case 1601–10).²⁵ Hence, our approach has the merit of being feasible with the available tools and does not seem to be unduly deficient.

Keywords, words which are significantly more frequent in corpus A when compared to corpus B, will be calculated using the Log Ratio measure. In order to see the key discourses in the corpus through the century, each decade in

the century will be subject to pairwise comparison with the next decade – for example, 1600s to 1610s, 1610s to 1600s, 1610s to 1620s, 1620s to 1610s, etc.

Finally, a brief comment on notation. In this book we will, when mentioning a word, italicize the word in question. Where we are referring to the lemma of a word, that is, the citation form of the word, which would cover its various forms in practice, we put the word in small capitals. So if we discuss the use of the word ‘dog’ in an example we would write *dog*, if we wished to refer to the noun *dog* in any or all of its forms, we would write DOG.

1.7 Conclusion

This chapter has introduced the aims of this book as well as problems with existing corpora and culturomics. We have also introduced the corpus we will use in the book and have discussed a series of approaches that we will take to the data, based principally on frequency, collocation and keywords. In the next chapter, we will switch to setting the stage for our analysis by providing a background on the history of prostitution in England. The goal of this chapter will be to outline the historical background of prostitution in early modern England and the experiences and practices of prostitutes in the seventeenth century. Chapter 3 will go on to examine how prostitutes were constructed in popular discourse of the time. We will take a deliberately broad approach in these chapters because, first, we do not assume any prior specialized historical knowledge on the part of our readers and, secondly, the analysis in this book is primarily corpus-based and will take place in the chapters which follow. While the legislative context is national, much of the detail in the chapter that follows relates to London. We know most about prostitution in that city and, given its status as a major centre of population, such a focus seems inevitable. We set aside for the moment the question of whether public discourse focused on London when discussing prostitution, though we do return to that question in Chapter 6. For now, we note that the focus of the review on London happily matches a bias we may expect in the data – most published works were produced in London. Hence, we might expect that public discourse as represented in EEBO will have a London bias. Given that this bias is in favour of a city where we also have most evidence of prostitution, the bias is a fortuitous one for the purposes of this book.

A second issue to note is that Chapters 2 and 3 focus on the practice of female prostitution in seventeenth-century England. Despite evidence indicating the existence of male prostitution in the same period being scant and ambiguous

to say the least, some scholars do believe that gay brothels were operating in London by the seventeenth century.²⁶ We intend to investigate the lexis of male prostitution – for instance, terms such as *stallion*, *he-whore*, *franion* and *spintry* – in a separate study elsewhere. However, in Chapter 5, Section 5.4.3 we do note that early modern nicknames used for prostitutes tended to be presented in a feminized way.

The main goal of this book is looking at representation, so before we can explore the lexis of prostitution and show how collocation can help us to reveal changing meaning attached to the words, and hence perhaps changing representation, we need to explore what evidence we have of attitudes towards prostitution in the period already. That will be presented in Chapter 3. Chapters 2 and 3 will provide us with the context in which our corpus-based investigation can be interpreted. This will then allow us, in Chapters 4 onwards, to carry out the investigation of HARLOT, JILT, PROSTITUTE, STRUMPET and WHORE in the context that Chapters 2 and 3 have painted.

Historical context is essential for this book – as should be apparent from the discussion so far. Without understanding the societies in which these words were produced in our corpus, it is very difficult to understand their use and meaning. Similarly, without a way of beginning to approach the use and meaning of a word in historical context, it is difficult to use documentary evidence to reflect upon that period. The approach taken in this book is, therefore, similar to that taken by McEnery (2005) and in a number of multidisciplinary studies by Wodak, notably, Heer et al. (2008). The work in this book can be viewed as a continuation of the discourse historical approach of these works which has been championed by Wodak. As this chapter began to elaborate the methodological context in which the study presented in this book is framed, so the following chapter begins the process of outlining the socio-historical context in which any corpus-based findings should be interpreted.