1 Introduction

The Special Relationship

The relationship between Britain and America could be said to have truly commenced on 4 May 1607, when the Virginia Company of London established Jamestown, the first permanent English settlement in what was initially referred to as the New World.[1] Jamestown was the capital of the colony from 1616 until 1699, although it is often overshadowed by the establishment in 1620 of the Plymouth colony in Massachusetts, associated with the Pilgrim Fathers. By the 1770s, mass emigration from Europe had resulted in around 2.5 million people living in America. Many of these citizens of the New World were keen to separate ties from Britain and escape the imposition of new taxes which were seen as unconstitutional. The British government responded by closing the port of Boston, and by April 1775, British and American forces had clashed at Lexington and Concord. Thirteen American colonies united to form a congress, which declared independence from Britain, and on 4 July 1776, the United States of America was founded.

Despite the rather acrimonious path to independence, the relationship between the UK and the USA has been consistently friendly ever since. The two countries have remained allies in a number of major conflicts, including World Wars I and II, the Korean War, the Cold War, the Gulf War, the Iraq War and the Afghanistan War. In 1946, after the end of the Second World War, British Prime Minister Winston Churchill made reference to the 'special relationship' between the two countries, which has since encompassed the sharing of intelligence, economic investment and policy and trends in fashion and music.

While English has been the dominant language of both countries, commentators on both sides of the Atlantic have, sometimes humorously, emphasised cultural (Kirk 2005; Kaufman and Macpherson 2004) and linguistic (Allwood

[1] The New World was generally used to refer to the Americas, consisting of the land masses of North and South America and various islands linking them. In this book, when I refer to America or American English, I am normally referring to what is known today as the United States of America.

1964; Alego 2006) differences. In a short story called *The Canterville Ghost* (1887), Oscar Wilde wrote, 'We have really everything in common with America nowadays, except, of course, the language', while *The Oxford Dictionary of Quotations* (Partington 1992: 638) claims that George Bernard Shaw referred to 'two countries divided by a common language'. As both American English and British English are broadly intelligible, it makes sense to view them as *varieties* of the same language rather than as separate languages, while also acknowledging that there is considerable variation within each variety (e.g. British English contains dialects like Geordie, Scouse and Cockney, each with distinctive words, pronunciations and ways of expressing grammatical relationships). For the purposes of this book, however, I am more interested in general differences between the two major varieties, American and British English. But I also want to consider an additional dimension: time. The study of language can be synchronic, comparing two or more varieties at a given point in time, or it can be diachronic, looking at how a variety changes over time. This book combines synchronic and diachronic analyses, comparing changes over time in both American and British English in order to be able to describe the major trends in language use in recent decades. Rather than simply taking a single time point which gives a snapshot of the state of the two language varieties, my analysis intends to consider the direction that the varieties are moving in – if there are particular differences, then have these differences become more pronounced over time, or are they narrowing? Or is one variety moving in the direction of the other?

Academic research on the English language has recently pointed to the increasing dominance of American English, linked to America's status as the only world superpower as well as its prolific cultural output and influence over the last century. For example, Leech's (2004) study of use of modal verbs (verbs which signify possibility or permission, like *should*, *must* and *could*) points to a 'British lag', indicating that (in use of such verbs at least) British English appears to be about 30 years behind trends in American English. Similarly, McEnery and Xiao (2005) have found further evidence for the British lag with regard to whether people choose to use the full or bare infinitive (e.g. the distinction between *help them to feel good* and *help them feel good*). However, Hebblethwaite (2012), writing for the BBC News website in September 2012, claims that due to the US screening of British television programmes like *Doctor Who* and *Downton Abbey*, along with increasing numbers of British journalists working in America, a number of 'Britishisms' are finding their way into US English. Examples are charted in a blog[2] by Ben Yagoda, who uses data from Google Ngrams, citing, for example, words and phrases like *poo*, *ginger*, *turn up*, *knock-on effects*, *keen on*, *chat up* and *sell-by date*. Yet, using

[2] http://britishisms.wordpress.com/.

a technique called the Manhattan Distance, Patrick Juola (2012) argues that the two language varieties have actually become increasingly lexically *dissimilar* over the last 100 years. More cautiously, Finegan (2004: 36) has argued that 'no one can confidently predict degrees of divergence or convergence between AmE and BrE in the future'.

Bearing in mind such a range of different claims, this book aims to address the following questions: to what extent are British and American English different, in what ways, and how have these differences altered over the last 100 years? In order to answer these questions, I utilise a method that has become increasingly popular in language analysis in recent decades, called corpus linguistics.

Corpus Linguistics

Corpus linguistics is largely a method or set of techniques which can be used to analyse language in use. Based on the principle of sampling, analyses are carried out on a carefully chosen selection of texts containing naturally occurring language so that generalisations can be confidently made about the variety that they came from. A collection of such texts is called a *corpus* (from the Latin word for 'body', plural *corpora*). The idea of samples of texts may imply that only a small amount of data is actually examined, whereas in actuality, many of these collections contain millions or even billions of words. As such samples are therefore too large for analysts to make sense of them by reading them all from beginning to end, computer software is employed in order to count linguistic phenomena, carry out statistical tests, sort the data and present them visually to humans so they can interpret them more easily. The computer tools aid analysis but do not actually constitute an analysis in themselves; it is only with human input and interpretation that the patterns identified by computers can be explained.

Many of the texts in corpora contain additional levels of information that have been added to them, either by humans or computer software or a combination of both. For example, when a computer program counts the words in a corpus, we may want to be able to distinguish between homographs, words which are spelled the same but have a different grammatical class or meaning (consider how *set* can be a noun, adjective or verb or can refer to a badger's home, a tennis score, a collection of musical pieces, a place where movies are made or how someone's mouth looks). If all of the words in a corpus are assigned codes which indicate this information, we can make more sophisticated and fine-grained calculations on the data. Particularly in the later chapters of this book, I make use of versions of corpora where words have been assigned grammatical or semantic codes.

A useful distinction is made within corpus linguistics between two types of research: corpus based and corpus driven (Tognini-Bonelli 2001). Corpus-based studies involve forming and testing hypotheses about language. These hypotheses may arise in a number of ways. For example, they may be based on a claim or finding made by someone else and can often be found through carrying out targeted literature review searches or by reading around a subject more generally. They can come about as the result of smaller-scale qualitative and/or quantitative analyses, often involving a pilot study or related data set. Additionally, corpus-based research can be serendipitous, involving a 'noticing' of a particular phenomenon in language as a result of our everyday encounters. We may then be motivated to determine whether an interesting feature of language is actually as widespread or becoming as popular as we think. But whatever the origin of the hypothesis, the researcher will know what he or she wants to look for in advance of approaching the corpus and will usually have a particular question in mind, such as 'Are nouns more common than verbs in recent American English?'[3] A potential limitation of this kind of research is that it requires humans to form hypotheses about what they think might be interesting about language, based on what somebody has noticed. Unfortunately, such an approach can be problematic, as we are burdened with numerous cognitive biases. For example, people tend to focus more on information that is encountered at the beginning of an activity (a cognitive bias known as the primacy effect; Murdock 1962), and we often discredit evidence which discounts our beliefs (the confirmation bias; Watson 1968). We also have a tendency to overestimate the importance of small runs or clusters in large samples of random data (the clustering illusion; Gilovich et al. 1985), and we have greater recall of negative events compared to positive events (the negativity bias; Kanouse and Hanson 1972). Hence, computer software, unhampered by such biases, is useful at objectively identifying the main trends and patterns. This ensures that nothing is overlooked and that we are able to hone in on features that we may not have considered ourselves. This kind of approach is termed a corpus-driven analysis; we begin the analysis from a relatively naïve perspective with no initial hypotheses. Instead, we may ask open questions, such as 'What characterises the language in this corpus?' or 'What aspects of language are different and similar in two corpora?'

One such corpus-driven technique is referred to as a keyword analysis. For our purposes, this involves comparing frequencies of all of the words in two corpora and running statistical tests to identify which words are much more frequent in one of the corpora compared against the other (the words which emerge in this way are referred to as keywords, described in more detail below).

[3] The answer is yes. In the 1 million word corpus of American English from 2006 that I am using in this book, there are 277,513 nouns and 178,687 verbs.

There are potential issues around corpus-driven approaches like keywords too. The first is that they often give too many results. As such approaches consider every word (or linguistic) feature in a corpus, the analysis will present information about each one, running into hundreds or thousands of rows of data. We usually want to focus on a subset of cases where the patterns of variation or change are most dramatic, so this means imposing cut-offs. Some corpus-driven methods involve carrying out statistical tests like chi square or log-likelihood, which elicit a *p* value, indicating the likelihood that we would have obtained the results we found if there were no change or difference between the frequencies of the feature in the corpus or corpora we are examining. However, such tests were not always designed with linguistic analysis in mind, and so using traditional cut-offs can still give hundreds of 'statistically significant' results. Another option is to use cut-offs based on rank orders of the statistical output, e.g. taking the top 10, 20 or 100 features that have the highest log-likelihood scores. This method at least produces a smaller set of features to focus on, although it should be borne in mind that such cut-offs are arbitrary, and thus our discussion of results will be based on how feature *x* shows comparatively more change over time in relation to features *y* and *z*, which appear lower down the list.

A second issue with corpus-driven analyses is that they can often tell us what we already know or would expect to find (although we should bear in mind another bias called the hindsight bias, also known as the 'I-knew-it-all-along effect'; Fischhoff and Beyth 1975). For example, it is hardly groundbreaking that a keyword analysis comparing American and British corpora would yield words like *color* and *colour* in each corpus, respectively. Very few people would be surprised to be told that a main difference between the two varieties is due to how certain words are spelled. For completeness, we may want to report what I have referred to as 'so what' findings (Baker and McEnery 2015: 8) but not spend too long on them, instead concentrating on those which are less expected. However, even obvious differences can sometimes inspire interesting questions. With spelling, for example, while it is obvious that there are differences between British and American English, what may not be so apparent is whether the differences are being steadily maintained over time or whether one variety is moving closer towards the other. As far as possible, I have tried to incorporate corpus-driven analyses into this book, although it is important to bear in mind that a hard distinction between corpus based and corpus driven is somewhat simplistic (McEnery and Hardie 2012: 143), and most research falls on a cline between the two.

In order to obtain a full and accurate picture of language change and variation in British and American English, in this book, I analyse a matched set of eight corpora encompassing texts of written standard published English. The coming chapters focus on different levels of language: orthography (Chapter 2),

affixation/letter sequences (Chapter 3), words and word sequences (Chapters 4 and 5), parts of speech (Chapter 6), semantics/culture (Chapter 7) and identity/discourse markers (Chapter 8). As well as reporting quantitative findings, the book goes beyond tables of figures and graphs by qualitatively examining cases of English in use and attempting to relate change and variation to social and historical context in order to interpret and explain findings. The following section introduces the corpora I will be working with.

Meet the Brown Family

The corpora that are used in this book compose a set of eight that are collectively known as the Brown Family. They are referred to as a family because they were all built using the same sampling frame, giving comparisons between them a high validity. Work on the first member of the family began at Brown University in the early 1960s, where W. Nelson Francis and Henry Kučera built what they called *A Standard Corpus of Present-Day Edited American English for Use with Digital Computers* but was later shortened to the Brown Corpus (demonstrating the trend of language densification, which will be encountered at various points in this book). The Brown Corpus consists of 1 million words of written standard English that was published in 1961. It contains samples[4] from 500 different text sources of about 2000 words each. Francis and Kučera (1979) wrote in the *Brown Corpus Manual* that 'samples were chosen for their representative quality rather than for any subjectively determined excellence. The use of the word *standard* in the title of the Corpus does not in any way mean that it is put forward as "standard English"; it merely expresses the hope that this corpus will be used for comparative studies where it is important to use the same body of data'.

The 500 text samples were taken from four main categories of writing (press, general prose, learned writing and fiction), which were further split into 15 subcategories or genres, labelled with the letters A–R (letters I, O and Q were not used). Table 1.1 gives a breakdown of the categories, along with the numbers of texts sampled in each. The texts were taken from the library at Brown University as well as the Providence Athenaeum and the New York Public Library (which kept microfilm files of the press articles used). Francis and Kučera (1979) describe how the categories and numbers of texts were decided by members of a conference at Brown University in February 1963. This included both Francis and Kučera as well as John B. Carroll, Philip B. Gove, Patricia O'Connor and Randolph Quirk. The numbers of texts in each genre are not equal but reflect

[4] Most of the samples did not consist of full texts but rather were 2000-word excerpts of longer texts. An exception is for newspaper articles, which are sometimes quite short, so several articles from the same newspaper were taken to represent a single 'text'.

Table 1.1 *Text categories in the Brown family*

Broad text category	Text category letter and description ('genre')		Number of texts: American corpora	Number of texts: British corpora
Press	A	Press: Reportage	44	44
	B	Press: Editorial	27	27
	C	Press: Reviews	17	17
General prose	D	Religion	17	17
	E	Skills, Trades and Hobbies	36	38
	F	Popular Lore	48	44
	G	Belles Lettres, Biographies, Essays	75	77
	H	Miscellaneous: Government documents, industrial reports etc.	30	30
Learned writing	J	Academic prose in various disciplines	80	80
Fiction	K	General Fiction	29	29
	L	Mystery and Detective Fiction	24	24
	M	Science Fiction	6	6
	N	Adventure and Western	29	29
	P	Romance and Love story	29	29
	R	Humour	9	9

what the linguists felt would be the most representative coverage of English writing. Random number tables were used in order to decide which texts to sample. The corpus was first published in 1964.

In the early 1970s, a second corpus was created, using equivalent texts from 1961, although rather than comprising American English, it was made of British writing. This corpus was created by collaborators at the University of Lancaster, the University of Oslo and the Norwegian Computing Centre for the Humanities at Bergen and so was known as the Lancaster-Oslo/Bergen, or LOB, corpus. The only difference in the sampling frame was to do with the numbers of texts collected in categories E, F and G, where there are slight differences, although as these categories tend to be somewhat more loosely defined and overlap more with each other than some of the others, this decision should not be seen as making comparisons between Brown and LOB invalid.

Since the publication of these first two corpora, six others have joined them. Christian Mair (1997) began an initiative to create matched corpora of the early 1990s, resulting in the production of the Freiburg-LOB Corpus of British English (FLOB) and the Freiburg-Brown Corpus of American English (FROWN). FLOB contained texts from 1991, while FROWN's texts were

Table 1.2 *The Brown family*

B-BROWN American English 1931	BROWN American English 1961	FROWN American English 1992	AmE06 American English 2006
B-LOB British English 1931	LOB British English 1961	FLOB British English 1991	BE06 British English 2006

published in 1992.[5] In the late 2000s, two more versions of the corpora were created at Lancaster University. I collected the texts that made up the British English 2006 Corpus (BE06) (see Baker 2009), while Amanda Potts led a team to create the American English 2006 Corpus (AmE06) (see Potts and Baker 2012). Due to the wealth of available data now online, texts were sampled from online sources, with the proviso that they needed to have first been published in 'paper' format so that comparisons with the earlier forms of published writing in the 1960s and 1990s corpora would be valid. Finally, two further corpora were added to the family, acting as precursors to Brown and LOB, with data sampled from a few years either side of 1931. A team led by Marianne Hundt at the University of Zurich collected the Before-Brown (B-Brown) Corpus, while Geoffrey Leech at Lancaster oversaw the Before-LOB (B-LOB) Corpus (see Leech and Smith 2005). As some of the corpora are named with similarly sounding acronyms which are not intuitively descriptive of their contents, I have decided not to refer to them by their names throughout the book. It is asking rather a lot to expect readers to memorise the periods and regions that eight different corpus names stand for. So instead of writing 'the FROWN corpus', I usually refer to 'the 1992 American corpus'. For reference purposes, Table 1.2 shows the relationships between the eight corpora, which can be realised as a 4×2 grid, with the rows showing language variety and the columns showing time period.

An issue with using the same sampling frame to create new corpora is described by Oakes (2009) and Baker (2010a). While the sampling frame may have accurately represented the types of writing (and the relative frequencies of people engaged in producing or consuming the different types) when it was initially created for the context of 1961, trying to match the sampling frame for a different time period (or location) may result in the corpus builders not properly

[5] The fact that FROWN and FLOB consist of texts collected a year apart does not mean that they cannot be directly compared. It may be the case that the later corpus (FROWN) might refer to slightly different world events, and this is something to take into consideration when carrying out the analysis.

capturing the way that written English is used at that point. For example, Oakes (2009) argues that the 1960s could be considered as the 'heyday' of science fiction writing, with a large number of science fiction books being published and read (relative to later decades). This would justify the inclusion of science fiction as a genre in the corpus. But if people were not reading as much science fiction in later decades, should we include the same number of such texts in a later corpus? And what about newer or emerging genres? For example, genres of fiction such as horror were not included in the Brown sampling frame, although it might be argued that by 2006, horror fiction was popular enough to warrant a section. A similar problem involves Category N, Adventure and Western. In the Brown Corpus, this included 'western' fiction, although while there have been British writers of western fiction (such as John Russell Fearn and Jim Bowie), focus in the British corpora was instead placed on adventure due to the fact that westerns are set in the American West. However, the lack of any category which properly matches American western fiction in the British corpora could be viewed as potentially problematic.

A possible solution to the fact that different cultures and time periods reflect interests in different genres is to try to use different categories from the original sampling frame, although it could be argued that this would make subsequent comparisons less valid. I thus acknowledge that the sampling frame for the Brown family is mostly static, and so findings and claims need to be restricted to the registers under examination. However, I feel that the benefits of keeping to the frame outweigh the disadvantages – this is an issue I return to in the concluding chapter.

Another point worth considering relates to the fact that all the samples are taken from *published* texts. They represent a somewhat 'conservative' form of English that is likely to have been subjected to proofreading and post-editing conventions to ensure it keeps within expected standards. However, a lot of the innovation in English happens in much more informal contexts, especially where young people or people from different backgrounds mix together (e.g. see Eckert 2003 or Torgersen et al. 2006). By the time such innovation finds its way into written published standard English, it is probably no longer innovative. So the Brown family is unlikely to be able to tell us about what is happening at the forefront of linguistic change. However, any changes that are noticed are likely to have already become well established, again meaning that findings have strong validity, even if they do not offer a great deal of insight into the newest uses of language.

Gathering a collection of 1 million words of language data was impressive in 1961 but by recent standards, the Brown family are now 'small' corpora. The British National Corpus (Aston and Burnard 1998), collected in the early 1990s, is 100 times larger than the Brown corpus, whereas the ukWaC corpus contains 2 billion words of online data, gathered from pages ending in the

domain.uk (Baroni et al. 2009). There are clear advantages to having corpora consisting of larger sample sizes; we can be more certain that our findings can be generalised to a population of language users, and we are more likely to find uses of relatively rare words, enabling us to include them in our analyses. But how much data would we need in order to be able to reach conclusions about different aspects of language? Kennedy (1998: 68) suggests that half a million words would suffice for an analysis of verb-form morphology but that a million words word not be adequate for a lexicographical study as up to half the words in such a corpus would occur only once. The study of grammar might require fewer words, however, as grammatical patterns tend to be more repetitive and there are a smaller set of grammatical categories. Biber (1993) has suggested that a million words would be enough for such studies. For the aims of this book, I argue that corpora consisting of 1 million words are large enough to focus on the phenomena that I am most interested in. The aim is to provide coverage of the most noticeable and oft-encountered differences and changes in English. So whether I look at lexis or grammar or some other feature, I will usually be focussing on patterns of change and stability around the most *frequent* features in a particular category. This allows justice to be done to a smaller number of features rather than presenting me with the unwieldy task of summarising patterns around every word in American and British English. For this reason, I have imposed quite harsh cut-off points for frequency phenomena, which has reduced the analysis to a manageable amount. In any case, a lot of language use can be accounted for by a small number of very frequent words. For example, for the four British corpora, the most frequent 380 words across them account for 62% of their total linguistic content (Baker 2011: 70), so an analysis which takes into account only these 380 words will tell readers about the language they are most likely to encounter. In general, I have considered features (words, tags etc.) which occur 1000 times or more across either the four British or four American corpora (e.g. on average 250 times per corpus).

Over the decades, the Brown family have enabled a great deal of corpus linguistic research, including much comparative research on grammar: prepositions (Lovejoy 1995), *do* as pro-form (Meyer 1995), progressive verbs (Smith 2002), modal verbs (Leech 2004), zero and full uses of infinitive marker *to* (McEnery and Xiao 2005). A particularly thorough comparison of the 1961 and 1991/2 members of Brown family is by Leech et al. (2009) which covers the subjunctive mood, modal auxiliaries, semi-modals, the progressive, the passive voice, expanded predicates, non-finite clauses and noun phrases. In devoting a single chapter to grammatical change as opposed to an entire book, I cannot hope to provide the same level of detail as Leech et al. (2009), but instead aim to (1) corroborate (or not) some of their findings by expanding their analyses to include a further four corpora, and (2) to identify in a more corpus-driven