Using Corpora in Sociolinguistic Research

Framing sociolinguistics within corpus research

In this chapter we consider how corpora can be used in order to carry out research from a sociolinguistic perspective. Sociolinguistics is a somewhat broad term, with Labov (1972: 183) indicating that it can appear redundant as all language is social. Despite this, sometimes language can be considered from less 'social' perspectives e.g. in terms of a description of how it is structured or to the extent that it resembles other languages. Researchers like Wardhaugh (2005) and Bloome and Greene (2002) have identified sociolinguistics as involving consideration of social context and the relationship between language and society. Sociolinguistics can thus involve analysing aspects of language use as they relate to a person's identity or the community they belong to. It can involve examination of variation between individuals or groups and change over time. In particular, it can concern interactive aspects of language, resulting in corpus-based sociolinguistic studies that have involved concepts like discourse features (Aijmer 2015), politeness (Culpeper and Gillings 2018), and representation of identity (Johnson and Partington 2017).

For the purposes of this chapter, we are taking a somewhat narrower focus, considering sociolinguistics from the perspective of speaker identity. We begin by reflecting on why a corpus approach is worth taking in order to answer sociolingusitic questions, followed by a short review of corpus studies of sociolinguistics. We then move on to a case study which is situated as a continuation of research on diachronic change in modal verb usage (e.g. Leech 2002, 2011; Millar 2009). While these studies have used written corpora, we have employed two spoken corpora in order to examine use of the modal verb *may*. The first is the spoken section of the British National Corpus (Aston and Burnard 1998), a general corpus of speech and writing collected between 1991 and 1994 (henceforth referred to as the BNC1994), which we compare against the Spoken British National Corpus 2014, a second corpus of speech from 2014 (Love et al. 2017), referred to as the BNC2014.

Great challenges bring great benefits

One consideration to be taken into account when using corpora to address sociolinguistic questions is that it tends to be easier to build or find corpora that consist of written texts, particularly if such texts are available in online contexts that already exist in electronic form (as opposed to say, converting the contents of a hand-written diary). Written corpora *can* be

used to examine sociolinguistic variation, particularly if records are kept regarding aspects of the identities of their authors. For example, Cermakova and Farova (2017) have examined variation between male and female authors in corpora of British and Czech fiction. However, many sociolinguistic studies tend to focus on naturally occurring speech, which has traditionally been more problematic for corpus linguists to examine. It can be complicated, expensive, and time consuming to collect the large amounts of data required to carry out a meaningful spoken corpus analysis. While the means of recording a conversation has become easier, thanks to the ubiquity of smartphones that can record sound, the resulting audio files still have to be keyed in using a consistent transcription scheme for representing accents as well as para-linguistic and non-linguistic features. Additionally, issues involving ethics including permission and anonymisation are if anything, more salient than they perhaps were thirty years ago. Additionally, corpus builders might struggle to construct a well-balanced corpus that contains speech from a wide range of speaker identities. The BNC2014, for example, contains speech from 671 people, and contains a wide range of identity categorisations, with some categories being better populated than others. If we start to combine categories (e.g. counting the possible permutations of age plus gender plus social class), some of the resulting demographic groups will contain very low numbers of speakers. Such studies can be easier if carried out on data collected online, however, e.g. Subtirelu's (2017) examination of the intersection of gender and race in student evaluations of lecturers on a 'rate my professor' website.

With those points considered though, the benefits of using a corpus (spoken or not) to investigate sociolinguistic issues are great and thus worth engaging with. A large amount of transcribed speech, encoded for different speaker characteristics, presents a huge advantage for analysts, both in terms of claiming representativeness about a particular social group and in terms of allowing both existing hypotheses to be explored. Furthermore, this method allows for the identification of linguistic features that may not have been noticed if the analyst had used manual means of identification or was limited to a much smaller dataset. For example, taking a keywords approach to compare different types of women's personal adverts, it was found that women who were seeking relationships with men tended to use the word *me* more often than those seeking relationships with women, resulting in the former group of women making higher use of statements where they positioned themselves as the recipient of actions of the desired other e.g. 'I just want someone who can make me happy' (Baker 2017).

Broadly speaking, corpora have been used in order to examine demographic variation, particularly utilising existing spoken corpora like the BNC1994 (see for example, Schmid's (2003) comparison of lexical items traditionally associated with male and female speech or Rayson et al's (1997) keyword comparisons of demographic differences. Some studies have used corpora as part of a larger multi-method study such as McEnery's (2005) examination of swearing in the BNC1994 which complemented historical analysis of texts about swearing in seventeenth and twentieth century written texts. While lexical and/or grammatical studies have generally been easier to carry out, due to the lower requirements in terms of annotation, there have also been (smaller-scale) studies that have considered phonetic or prosodic variation such as Grabe and Post (2002) examination of stress in different dialects or Maclagan and Hay (2007) who looked at pronunciation in a corpus of New Zealand speech, and use of short monophthongs by younger and older speakers in southeast England, Torgersen et al. (2006).

Another issue with spoken corpus projects is that they tend to be based on collecting data at a single point in time, so it can be difficult to replicate the conditions of an earlier spoken corpus building study in order to examine change over time. Diachronic studies have again tended to involve written corpora, bringing us to a point which this chapter aims to address.

Changing frequencies of modal verbs over time

The case study we describe in the following section is based on comparing spoken corpora to identify change in modal verb use over time. Modal verbs are verbs of necessity or possibility like *would*, *should* or *may* which are often used in conjunction with the base form of another verb (e.g. *might go*). In a diachronic study which compared four British and American reference corpora (known as the Brown Family) containing written texts taken from two time periods, Leech (2002) found that the collective frequencies of eleven modal verbs were lower in the 1990s corpora, compared to the 1960s ones. This pattern was more pronounced for the American corpora, which also had lower frequencies of the modal verbs in both time periods, when compared to the equivalent British corpora. In particular, verbs which expressed strong modality like *shall*, *need*, *must* and *ought* showed much lower differences in frequency between the two time periods, compared to weaker modals like *might*, *will*, *would*, *can* and *could* (with the latter two modals being slightly more frequent in the 1990s American corpus, compared to the 1960s one). On the other hand, a group Leech refers to as semi-modals, like *had better*, *want to* and *have (got) to* had higher frequencies in both 1990s corpora, and had

higher use overall in the American corpora, compared to the British ones. Leech suggests that the patterns reflect a wider trend towards colloquialization of written language, shown by adoption of features more common to speech (contractions, progressive verbs, questions, genitives, zero relative use), appearing more often in writing. He cautiously discusses possible other social explanations like democratization (a tendency to avoid unequal and face-threatening modes of interaction) and Americanization (the influence of North American habits of expression and behaviour on other nations).

A related corpus study of modal verbs (Millar 2009), analysed the 100 million word TIME magazine corpus, containing text from issues of the American *Time* magazine from 1923 onwards. The study found a 22% increase in modals overall, between the 1920s and 2000s, although some modals like *shall* had drastically decreased in frequent over time, some, like *would* showed fluctuation while others like *may*, *can* and *could* had increased. Leech (2011) responded by arguing that the language of *Time* magazine cannot be seen as representative of all language use, and by including the analysis of additional members of the Brown family of corpora, he shows that modal verb decline is most notable when comparing corpora from the 1960s, 1990s and 2000s. Similarly, his analysis of the Corpus of Historical American English (COHA) shows decline in modal use across the decades of the 20th century which accelerates from the 1970s. His study notes that *may* is a bone of contention in that it is where his findings disagree most dramatically from Millar (2009). While *may* appears to have decreased over time in the COHA, during the same period it has increased over time in the TIME corpus. Both Millar and Leech suggest that the rise in frequency of *may* in Time magazine could be due to change in content, style or editorial policy, with a shift towards speculation in reporting and less focus, particularly in magazines, on the past and present.

In British English, *may* has been associated with politely powerful ways of speaking. Stubbs (1996) describes how the word was used frequently in a court case by a judge during his summing up, in phrases like ‘you may think/feel/remember’ that were addressed to the jury. In some cases, the judge used *may* to signify a course of action he intended to take e.g. ‘when I sum up I may very well make some comments upon the evidence’, while in others the judge appeared to be politely instructing the jury e.g. ‘one of the questions you may want to ask yourselves is this’. In the spoken section of the 1994 British National Corpus, *may* is used more frequently by AB speakers (the highest social class), although this is not the case for the spoken BNC2014 corpus. The creation of a second British National Corpus, however, opens

up numerous opportunities for the investigation of diachronic variation, as we will demonstrate in the following section.

Case study: spoken use of the modal verb *may*

In this chapter, we examine diachronic changes in how different age groups use the modal verb *may*. In order to do this, we use two comparable corpora of speech (the spoken section of the BNC1994 and the BNC2014). The BNC1994 contains 100 million words, although only 12 million of these are transcribed speech. Within the transcribed speech, 7 million words originate from 'context governed' speech- such as language produced in the workplace. The remaining 5 million words derive from private conversations. The BNC2014 contains 11.4 million words of speech, all of which come from private conversations. For comparability, we have only used the 5-million-word sub-corpus of private conversations from the BNC1994 and the whole of the BNC2014. The two corpora are highly comparable and use the same age categorisation system: 0-14-year olds, 15-24-year olds, 25-34-year olds, 35-44-year olds, 45-59-year olds, and 60+-year olds. Thus, we can study the diachronic change from two perspectives. First, we can compare how the same age groups use language in different time periods, for example, comparing one set of people who were 0-14-years old in the BNC1994 with a second set of people who were 0-14 years old in the BNC2014. This kind of analysis allows us to examine how frequently 0-14-year olds in both 1994 and 2014 used the modal verb *may*. Additionally, we examine differences over time within the groups of people who were born around the same time. If we assume that the demographically sampled speakers are representative of their cohorts at the time of recording, it would thus be possible to observe how four different group's use of the modal verb *may* changes as they age. These cohorts are presented below in Table 1.

Cohort birth year	Age range in BNC1994	Age range in BNC2014
People born in the 1950s	35-44	60+
People born in the 1960s	25-34	45-59

People born in the 1970s	15-24	35-44
People born in the 1980s	0-14	25-34

Table 1 Age cohorts and how old members of those cohorts would have been in the BNC1994 and BNC2014

Each of the age groups is representative of people born within decades, rather than across decades. That is to say that the cohorts will stretch from, for example, the birth years of 1950-1959 as opposed to, for example, 1954-1963. Throughout the rest of this case study, when we refer to age groups, we refer to set age categories (for example, 0-14 year olds). When we refer to decade of birth cohorts, we refer to the groups notes in Table 1 (for example, people born in the 1980s).

Any diachronic differences in how members of a given decade of birth cohort use language could reflect two types of change. The first type of change relates to what is seen as socially appropriate uses for a modal verb, such as how the modal *ought* is now seen as archaic. The second change relates to how a birth cohort view their use of the modal verb as they age (or how they believe they should be talking for their age). In other words, diachronic changes could reflect both what is viewed as an appropriate way to use a modal verb for the time and how age groups apply this to their own language, bearing in mind their age at a given point in time.

With regard to the present study, the approach to analysing decade of birth cohort changes in tandem with age category can reveal nuanced differences in social variation. For example, if 0-14-year olds in 1994 were taught that *may* is an impolite way to modalise a construction, but 0-14-year olds in 2014 were taught that it is polite, then it is possible that the latter group would use *may* more frequently, and this would demonstrate a diachronic change. However, if 25-34 year olds in BNC2014 use the modal verb in a way which is a) different to when they were 0-14, b) different to the 25-34 year olds from the BNC1994, and/or c) different to the 0-14 year olds in the BNC2014, it could point to the decade of birth being an important social variable. The two different approaches to age (age group and decade of birth cohort) form the basis of the analysis in this chapter. By taking into account both variables (age group and decade of birth cohort), we can obtain a better picture of variation in the use of

may among British English speakers at a diachronic level, something which until the availability of the BNC2014, has not been easy to do.

Both corpora are hosted on CQPWeb - a freely accessible website hosted at Lancaster University (Hardie 2012). CQPWeb has numerous functions which are indispensable to sociolinguists. In this study, we use the data disclosed by speakers to investigate how different social groups use language. All the speech in the corpora has been tagged for speaker, and when recording the data, those speakers disclosed their social identities. These identities include, but are not limited to, age, gender, and social class. These identities are included in the metadata for each speaker, meaning that analysts can look at how members who hold identity 'x' use feature 'y'. Additionally, we use the Distribution function in CQPWeb, which allows a researcher to examine how frequently members of a particular social identity use a specified word. While the Distribution function allows researchers to examine how frequently a search term is used, it does not necessarily explain how the term is used. To that end, we use the Restricted Query function for both corpora. Apart from the Restricted Query function allowing researchers to select sub-corpora in the BNC1994, it also allows analysts to examine all the occurrences of a search term when used by a specified social group. This can thus lead to more qualitative analyses of concordance lines, which in turn can reveal how the search phrase is used.

A final consideration in using these corpora is that CQPWeb automatically applies the CLAWS tagger to uploaded corpora. The CLAWS tagger annotates the grammatical part of speech for each word and is accurate 96-97% of the time (see Fligelstone et al., 1997). *May* is polysemous, meaning that if researchers do not specify the part of speech they are interested in, CQPWeb will return all forms of *may* - including not only the modal form but also forenames, surnames, and references to the month of May. The BNC1994 is tagged using the C5 tag set, and the BNC2014 is tagged using a modified version called the C6 tag set. There are minor differences between the two tag sets (for example, modal verbs are tagged as VM0 in the C5 tag set and as VM in the C6 tag set), but this does not impact on our results. When searching for a part of speech, a researcher would have to include an underscore to show that they only desire results of a particular grammatical category. To restrict the results of *may* to only show instances of it used as a modal verb in the BNC1994, this would involve the researcher using the search term 'may_VM0'. In the BNC2014, this would involve the researcher using the search term 'may_VM'.

Our analysis uses two different methods for interpreting the data: the first method is to analyse the changes in frequencies at which the modal verb *may* occur. The second method we implement is a close reading of extended concordance lines in order to examine if the function of the modal verb has diachronically changed. Within this latter method we also explore the phraseological patterns to examine how speakers use language to achieve the functions associated with the modal verb (requesting permission, hedging, and giving permission). For both methods, we explore differences by comparing age groups and cohorts. However, in the analysis derived from the second method, we focus on three groups in particular: those who were aged 0-14 in the BNC1994, those who were 0-14 in the BNC2014, and those who were 25-35 in the BNC2014 (members of this group would have been 0-14 in the BNC1994).

Before we examine how different age groups use *may*, it is first worth knowing how frequently it is used in the corpora as a whole. Without imposing the restriction of age categories, within the BNC1994, *may*_VM0 occurs 637 times across 97 texts (127.03 occurrences per million words). In the BNC2014, *may*_VM occurs 1,365 times across 592 texts (119.50 instances per million words). Thus, the general usage of *may* appears to have slightly decreased across time at a rate of 7.53 occurrences per million words. However, this does not reveal how different social groups are using it. For example, in the BNC1994, all age groups might use *may* with around equal frequency, while in the BNC2014 five of the six age groups in the data might have an extremely low use of *may*, with one age group using it at a much higher frequency. Figure 1 outlines the diachronic differences in the frequency at which *may* (as a modal) occurs across comparable age groups.

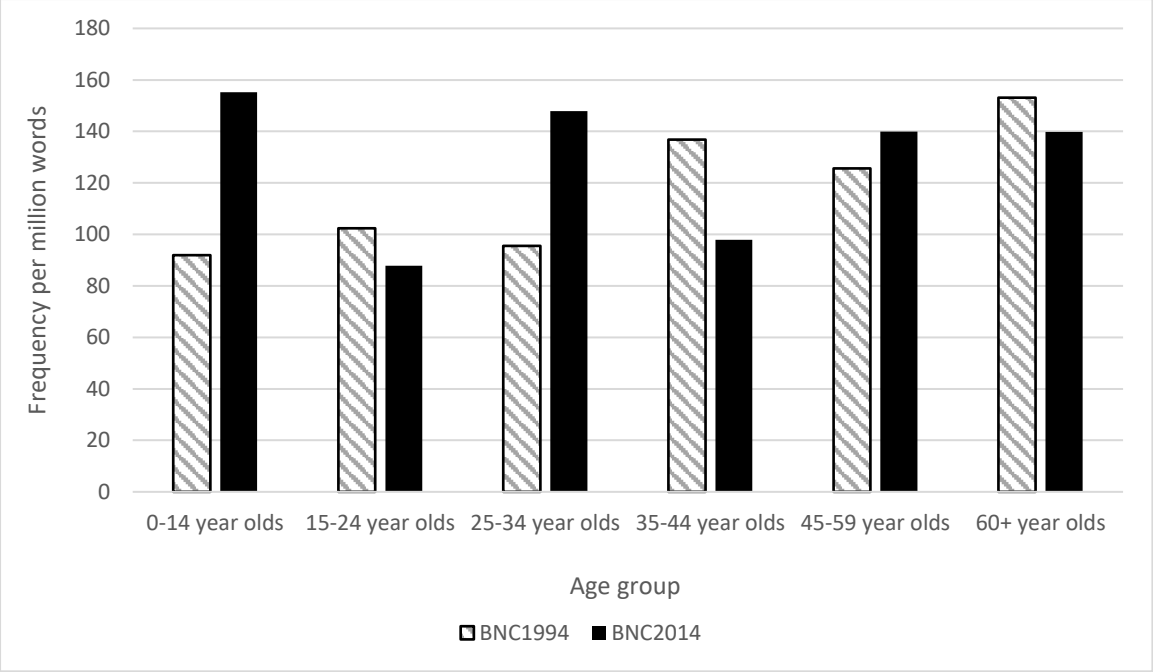

Figure 1 Diachronic change in *may* as a modal verb across comparable age groups.

The data from Figure 1 paints a rather complex picture: across all six comparable age groups, three (0-14-year olds, 25-34-year olds, and 45-59-year olds) appear to use the modal verb *may* more frequently in 2014 than in 1994. In contrast, the other three (15-24-year olds, 35-44-year olds, and 60+-year olds) appear to use *may* as a modal verb less frequently in 2014 than in 1994. These changes are important when we consider that the modal form of *may* was meant to be declining in frequency in a general sense. The frequency at which *may* occurs as a modal in the BNC2014 as a whole has decreased by 7.53 per million words. But when we examine how often different groups use it, the amount of change ranges from 0-14 year olds, who use *may* 63.36 times per million words more often in 2014 compared to 1994, to 35-44 year olds who use *may* 38.99 occurrences per million words less in 2014 compared to 1994. These differences in usage compared to the general picture of diachronic changes to the modal form of *may* give weight to why sociolinguistic analyses of corpora are needed. Although examinations of diachronic change in the frequency at which a word is used can provide interesting findings, they do not necessarily reflect the behaviours of different social groups, which can massively deviate from the perceived diachronic change in a general sense.

One pattern which appears to emerge is that alternate generations are using *may* more frequently than in 1994, while the remaining generations are using it less frequently than in

1994. Importantly, no age group in the BNC2014 used the modal verb at the same frequency as their comparable age group did in the BNC1994. Therefore, it is possible to suggest that these social groups may have different preferences to the use of *may* as a modal verb and as such are using it in either more/less frequently in the BNC2014 than in the BNC1994.

Although it is useful to know that there is a diachronic change in the frequency at which the modal verb *may* is used across comparable groups, it does not answer the question of whether a particular decade of birth cohort keeps these features as they age. Thus, these diachronic changes within decade of birth cohorts are presented below in Figure 2.

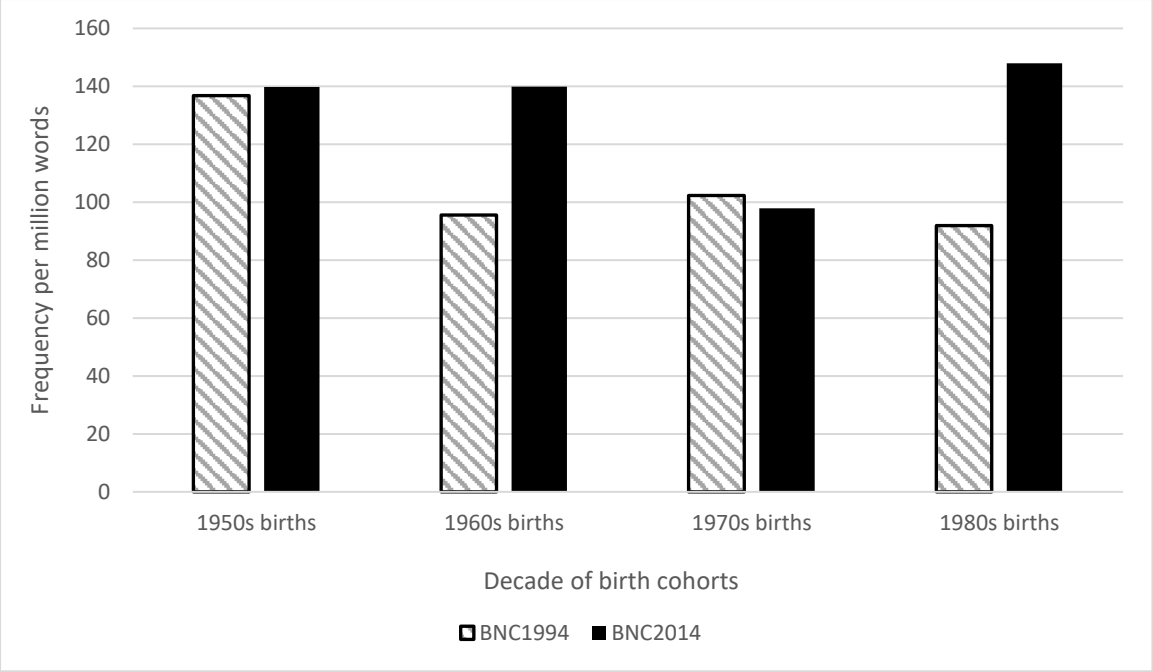

Figure 2 Diachronic changes in how frequently decade of birth cohorts used the modal verb *may*.

Figure 2 indicates a different picture compared to the data presented in Figure 1. With the exception of those born in the 1970s, every decade of birth cohort appears to have increased their frequency of usage of the modal verb *may*. When examining age group differences, the number of groups which increased/decreased in their usage of *may* was equal. However, there are more cohorts that increase in their usage of *may* as they age. This increase is most noticeable in those born in the 1980s who use *may* at a frequency of 56.05 instances per million words more frequent than they did in the BNC1994. Additionally, those born in the

1970s most closely reflect the general diachronic changes in the frequency at which *may* is used. This decade of birth cohort use *may* less frequently than they did when they were 15-24, at a rate of 4.49 occurrences per million words. In comparison to the smallest decrease in the analysis that only compared age groups in both corpora, the changes in this decade of birth cohort appear to more closely conform to the decrease of 7.53 instances per million words which occurs in diachronically in a general sense. Given that there are cohort differences which do not match the age group differences, it appears that both the time period a person is living in (whether it is the 1990s or 2010s) and their age at the time play a part in the frequency which they use *may*.

Quantitative findings such as these can provide a useful broad picture of diachronic changes in how a word is used. However, it is useful to combine them with an analysis of context, as this can reveal how a particular word is used in different time periods by different age groups and members of decade of birth cohorts. For example, it could be that in the BNC1994 most instances of *may* could be used to request permission to do something, while in the BNC2014, most instances could be used to hypothesise about the future.

In order to gain a clearer understanding of how *may* is used in both corpora, we conducted a close reading of concordance lines. When working with reasonably large frequencies, it can often be useful to examine collocates of a word (see for example Hunston, this volume; Jones, this volume), as a way of down-sampling concordance lines to a manageable amount. However, when we considered occurrences of *may* based on use within particular age groups, we found that its frequency was too low to produce collocates that met traditional significance thresholds (for a discussion of these thresholds see Brezina, 2018). Given the limitations of space, we only focus on three groups: those who were 0-14-years old in the BNC1994, those who were 0-14-years old in the BNC2014, and those who were 25-34 in the BNC2014. These groups were selected for closer analysis because 0-14-year olds in the BNC1994 and the BNC2014 had the greatest difference in normalised frequencies at which *may* occurs as a modal verb (see Figure 1). In other words, 0-14-year olds in the BNC2014 used the modal verb *may* much more than their comparable group from the BNC1994. Furthermore, the group who were 0-14-years old in the BNC1994 (who are the same decade of birth cohort as the 25-34-year old group in the BNC2014) had the largest increase in the normalised frequency of use of *may* as modal as they aged (see the 1980s births in Figure 2). We thus start by analysing how these age groups used the modal verbs in the BNC1994

compared to the BNC2014, in terms of whether they use it to request permission (e.g. *may I have some milk*), hedge propositions (e.g. *you may want to go*), give permission (e.g. *you may do that*), or other uses (e.g. inaccurate tagging). These quantified frequencies (as a percentage of the raw total for each group) are presented below in Figure [3]

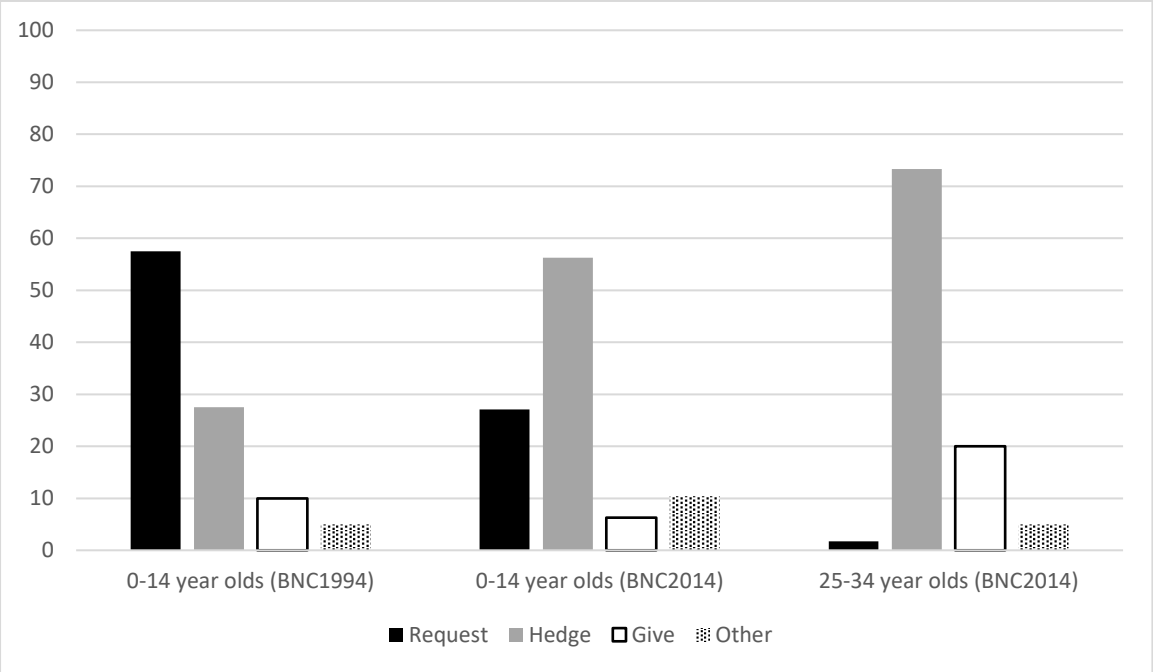

Figure 3 Percentages of the functions of *may* in three groups.

The above table suggests a startling change - both diachronically across the comparable age group and within the 1980s decade of birth cohort. First, across the comparable age groups (consider the first two sets of columns), the 0-14-year olds in the BNC2014 appear to use *may* as a modal to hedge more than their 1994 counterparts. Similarly, this change is coupled with a lower frequency at which *may* is used to request. Interestingly, across these comparable age groups, the use of *may* to grant permission appears to have only gone down slightly, though it is interesting to see that the frequency at which this occurs is lower than the 'Other' category in the 2014 data set, but it is not in the 1994 data set. Given how low the frequencies of these are, it appears as though 0-14 year olds do not regularly use this modal verb to grant permission in either data. In the same vein as age group comparisons, the data suggests that 25-34-year olds within the BNC2014 almost never use *may* to modalise a request (4% of occurrences). In comparison to 0-14-year olds within the same data set, 25-34-year olds were more likely to use this modal verb to suggest some form of hesitancy and

hedge propositions. Finally, by comparison to both younger age groups, they were more likely to use it to give permission to do something.

When comparing how members of the same decade of birth cohort (those born in the 1980s) used language 20 years after members of the same cohort were initially sampled, more changes can be seen. Comparing the first and third sets of columns, members of this cohort were most likely to use *may* to request permission when they were children (0-14 years old), but as adults aged 25-34, they had almost entirely dropped this use of *may*. Similarly, members of this cohort were twice as likely to use the modal form of *may* to suggest uncertainty and hedge propositions as adults in comparison to when they were 0-14 years old. Thus, not only are there diachronic differences in how comparable age groups use language, but there are diachronic differences in how members of the same decade of birth use language.

However, while quantifying how frequently these different functions occurs may be helpful in painting a picture of diachronic change on a quantitative level, it does not necessarily explain how the modal verb is used at a phraseological level. That is to say, it does not necessarily reveal how *may* is used to achieve these functions. In order to do this, we examined the phraseological patterns around *may* for these groups within the data sets.

Table 2 below notes the different phraseological lexical units in which *may* was used. In order to ascertain these, we used the 'sort' function in CQPWeb. This function orders concordance lines by alphabetising co-occurring words in a pre-set slot to the left or right. Although analysts have control of which slot is set for the sort function, the results we present are based on sorting the concordance lines by the co-occurring word one place to the right. We only noted a phraseological pattern if it occurred in at least 10% of the concordance lines for a particular group. This prevented low-frequency patterns being taken as representative usages.

Age group (Data set)	Raw frequency	*Notable phraseological lexical units* (raw frequency of occurrence) example (filename)

0-14-year olds (BNC1994)	40	*may I* (21) *may I have my pudding please?* (KBW 9002)
0-14-year olds (BNC2014)	48	*may be* (5) *it may be a chocolate goose it may be a giant Easter egg* (S8LG 404)
		may have (6) *she touched him and got it and er and she may have been electrocuted I can't remember* (SNZS 175)
		may I (12) *may I have some cake?* (SCG9 533)
		may not (8) *she thinks she's a witch but she may not be a witch* (S46J 244)
25-34-year olds (BNC2014)	240	*may as well* (24) *if there's a little two you can make you may as well try and make it* (S968 1083)
		may be (43) *United States somewhere that's bigger where if we get where we may be able to afford property* (SVZB 1386)
		may have (33) *he may have been drunk when they were talking about this deal* (S4S5 631)

Table 2 Common phraseological units containing may for different age groups.

One immediate difference between the language used by 0-14-year olds in the BNC1994 and the BNC2014 is that those in the BNC1994 appear use *may I* much more regularly than those in the BNC2014. In both data sets, the modal construction of *may I* is used in order to request permission to do something and to request physical items, which could explain the difference in the frequency at which *may* is used to request permission demonstrated in Figure 3. However, the 0-14 year olds who were recorded for the BNC2014 use a greater variety of

phraseological constructions which incorporate *may*, compared to children of a similar age who were recorded for the BNC1994. Additionally, the phraseological constructions appear to suggest that the 0-14-year olds in the BNC2014 use constructions containing the modal form of *may* to talk about events they remember from the past, request items, and hedge ideas (such as seeing if something is a goose or an egg). This vastly contrasts to the 0-14 year olds from the BNC1994, who would mostly only use it to request items. In turn, these phraseological patterns provide some qualitative explanation for the data presented in Figure 3.

In contrast to both 0-14-year-old groups, the 25-34-year old group in the BNC2014 appears to use *may* in ways which suggest uncertainty about more abstract concepts, such as finances and biological phenomena. Furthermore, the 0-14-year olds in the BNC2014 exclusively use *may have* to refer to past events. In the BNC1994, those who were 0-14 only used *may have* three times (so it was not considered a phraseological lexical unit): two of these occurrences were references to past events, and one was a request. Comparatively, the 25-34-year olds in the BNC2014 appear to use *may have* to refer to both past events and potential future events, for example: *we may not need the twelfth month we may have finished all our work and done everything and therefore that would save us like seven hundred and fifty quid in rent* (SBTC 1149). Indeed, for this group there are seven occurrences (out of 33) which use *may have* to refer to future possibilities, all of these refer to events in the distant future, rather than the immediate future. Thus, in the BNC2014, the way different age groups use the modal verb phrase *may have* shows a diachronic change. Thus, members of the decade of birth cohort who were born in the 1980s appear to have changed how they have used the modal verb phrase.

When specifically examining the difference between how members of the decade of birth cohort born in the 1980s use language, it can be seen that they no longer retain the frequent use of *may I.* This could reflect how politeness forms that are considered appropriate for 0-14 year olds might be discarded later in life.

Reflecting and future directions (1000)

This case study has demonstrated how corpora can be used to explore not just social variation but also diachronic changes with regard to the use of modal verbs. It is now worth stepping back and considering some of the issues with using corpora in this way. In particular, the

study above raises issues regarding essentialist approaches to language. Throughout, we have tried to avoid the claim that 'x' age group use 'y' feature because they are part of the 'x' age bracket. To exemplify this problem, one must ask 'when someone has a birthday which pushes them into the next age bracket, do they suddenly stop using a linguistic feature?' The likely answer to this is no. In turn, this serves to demonstrate the issues of using statically categorised social groups. Although we have done our best to avoid essentialist claims, the practice of putting people into categories and then counting frequencies of linguistic features does face the risk of appearing essentialist, particularly as we cannot account for every intersection of a speaker's identity (age, social class, gender), and even when we are able to account for a number, these identities may not even play a role in the language used by that speaker (for a discussion of intersectionality, see Crenshaw, 1990). Importantly, to avoid essentialist thinking, a sociolinguistic corpus analysis should first note that differences in frequency are usually not absolute but more a matter of gradience with one social group using a feature more than others, but often not having full ownership of that feature. To that end, the qualitative forms of analysis favoured by examining concordance lines allow us to consider differences within a particular social group as well as differences between them. Considering context of usage via an analysis of an expanded concordance line allows us to identify individual speaker variation that can indicate that people within a group can use a word for different purposes (and indeed, the same speaker can vary in the way they use a word). Such considerations will help us to provide a more sophisticated sociolinguistic analysis that goes beyond a table of numbers and claims that group 'x' use feature 'y'.

Given the limitations of space and the limitations of the corpora (which would need to be even bigger to downscale sample sizes to account for more identities), it has not been fully possible to explore all intersections of identity in our exploration. While it would be interesting to compare these changes to how other social groups use the same feature, and how adding other dimensions changes the variation, even larger corpora would be required to do so.

Future research that utilises corpus approaches to sociolinguistics may want to focus on using corpus methods on data generated from communities of practice (see Eckert and McConnell-Ginet, 2007). By using language from communities that are known to organise around a particular identity, it might be possible to compare and contrast variation which is community specific, and avoid some of the essentialism that naturally occurs when using large reference

corpora. Additionally, future research may want to examine both different modal verbs and how they are used within the social groups that are the focus of this chapter. It would be interesting, for example, to examine other modal verbs are used to construct politeness and compare this across age groups. Elsewhere, other researchers may want to examine how the modal verb *may* changes across different identities and how and why these identities might play a role in the use of the modal verb *may.*

Finally, this study has only been able to answer the question 'how do these groups use language?' Future research might elect to attempt to answer the question 'why do these groups use these features?' Although we have attempted to make a couple of plausible explanations, a deeper and more nuanced study, which examines attitudes towards these kinds of modal verbs, could complement this one, providing further insights into this kind of social variation.

Further Reading

Baker, P. (2010) *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

This book acts as a general primer for a range of ways that corpora can aids sociolinguistics, having chapters on demographic variation, comparing language use across different cultures and examining language change over time, studying transcripts of spoken interactions and identifying attitudes or discourses.

Friginal, E. (ed). (2017) *Studies in Corpus-based Sociolinguistics*. London: Routledge.

This edited collection of 14 chapters from a range of authors is divided into three sections (1) languages and dialects (2) social demographics (3) register characteristics.

Friginal, E. and Hardy, J. (2013) *Corpus-based Sociolinguistics: A guide for students*. London: Routledge.

This book functions as a practical guide for students who wish to carry out their own studies, containing case studies, discussion questions and activities.

Murphy, B. (2010) *Corpus and Sociolinguistics: Investigating age and gender in female talk*. Amsterdam: John Benjamins.

This monograph involves a detailed analysis of age and gender in a 90,000 word spoken corpus of Irish English, considering features like hedges, vagueness, intensifiers and swearing.

References

Aijmer, K. (2015) Analysing discourse markers in spoken corpora: *Actually* as a case study. In P. Baker and T. McEnery (eds) *Corpora and Discourse Studies*. London: Palgrave pp. 88-109.

Aston, G. and Burnard, L. (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baker, P. (2017) Corpora and social demographics. In Friginal, E. (ed). *Studies in Corpus-based Sociolinguistics*. London: Routledge, pp. 157-77.

Bloome, D. and Green, J. (2002) 'Directions in the sociolinguistic study of reading.' In P. D. Pearson, R. Barr, M. L. Kamil and P. Mosenthal (eds) *Handbook of Reading Research*. Mahwah, NJ: Lawrence Erlbaum, pp. 395-421.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Cermakova, A. & Farova, A. (2017) *His eyes narrowed — her eyes downcast:* contrastive corpus-stylistic analysis of female and male writing. *Linguistica Pragensia* 28 (2), 7-34.

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43* (6), pp. 1241-1299.

Culpeper, J. and Gillings, M. (2018) Politeness variation in England. A North-South divide? In Brezina, V., Love, R., and Aijmer, K. (eds) *Corpus Approaches to Contemporary British Speech*. London: Routledge, pp. 33-59.

Eckert, P. and McConnell-Ginet, S. (2007). Putting communities of practice in their place. *Gender & Language, 1*(1), pp. 27-37

Fligelstone, S., Pacey, M. and Rayson, P. (1997) How to generalize the task of annotation, in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 122-136.

Grabe, E. and Post, B. (2002) 'Intonational variation in English.' In B. Bel and I. Marlin (eds) *Proceedings of the Speech Prosody 2002 Conference*, 11-13 April 2002, Aix-en-Provence: Laboratoire Parole et Language, pp. 343-346.

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, *17*(3), pp.380-409.

Johnson, J. and Partington, A. (2017) Corpus-assisted discourse study of representations of the 'under-class' in the English-language press. In Friginal, E. (ed). *Studies in Corpus-based Sociolinguistics*. London: Routledge, pp. 293-318.

Labov, W. (1972) 'The logic of nonstandard English.' In P. Giglioli (ed.) *Language and Social Contex*t. Harmondsworth: Penguin, pp. 179-215.

Leech, G. (2002) 'Recent grammatical change in English: data, description, theory.' In K. Aijmer and B. Altenberg (eds) *Proceedings of the 2002 ICAME Conference*, Gothenburg, pp. 61-81.

Leech (2011) The modals ARE declining. Reply to Neil Millar's 'Modal verbs in TIME: Frequency changes 1923-2006'. *International Journal of Corpus Linguistics* 16(4): 547-564.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017 fc). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. In *International Journal of Corpus Linguistics*, 22(3): 319-344.

Maclagan, M. A. and Hay, J. (2007) 'Getting fed up with our feet: Contrast maintenance and the New Zealand English "short" front vowel shift.' *Language Variation and Change* 19, 1-25.

McEnery, T. (2005) *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London: Routledge.

Millar, N. (2009) Modal verbs in TIME. *International Journal of Corpus Linguistics* 14(2): 191-220.

Rayson, P., Leech, G., and Hodges, M. (1997) 'Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus.' *International Journal of Corpus Linguistics* 2:1, 133-152.

Schmid, H-J. (2003) 'Do men and women really live in different cultures? Evidence from the BNC.' In A.Wilson, R. Rayson and T. McEnery (eds) *Corpus Linguistics by the Lune*. *Lódź Studies in Language 8*. Frankfurt: Peter Lang, pp. 185-221.

Stubbs, M. (1996) *Texts and Corpus Analysis*. London: Blackwell.

Subtirelu, N. (2017) Exploring the intersection of gender and race in evaluations of mathematics instructors on Ratemyprofessors.com. In Friginal, E. (ed). *Studies in Corpus-based Sociolinguistics*. London: Routledge, pp. 219-235.

Torgersen, E., Kerswill, P. and Fox, S. (2006) 'Ethnicity as a source of changes in the London vowel system.' In F. Hinskens (ed.) *Language Variation - European Perspectives. Selected Papers from the Third International Conference on Language Variation in Europe* (ICLaVE3), Amsterdam, June 2005. Amsterdam: Benjamins, pp. 249-263.

Wardhaugh, R. (2005) *An Introduction to Sociolinguistics*. London: Blackwell.