with singular human reference (Chapter 9), its normalised frequency does not show a consistent increase as there is a sharp drop between the periods 1720–39 and 1740–59.

We could also compare these frequencies with those found in other sources. The eighteenth century is covered by such materials as the *Old Bailey Corpus* and the *Corpus of Late Modern English Texts*, and it would certainly be useful to complement our data with these larger datasets in the future. As a first step, the next section compares the results we have gained using CEECE with those observable in the massive Google Books database.

13.2 Google Books: A shortcut to studying language variability?

Mikko Laitinen and Tanja Säily

This section looks into evidence provided by the big but messy database of Google Books (Google 2013a), which extends diachronically to Early Modern English, covering also the time period examined in this book. The purpose is to explore what types of broad diachronic evidence related to grammatical variability can be obtained from this massive freely available database with respect to the changes studied in this volume, to complement the insights into sociolinguistic variability offered by our specialized and tailor-made corpus in Part II.

The motivation for this type of comparison of big and messy databases with small and tidy corpora comes from two sources. The first is related to Google itself and the effects its search engines are said to have caused on information seeking in general. Google Books is, just like many but not all of the tools provided by Google Inc., an easily available free tool that enables access to a vast amount of information. The success of Google has led to the coining of a popular phrase, the Google generation, which in popular myth refers to the generation born after 1993 and is used to refer to a generation whose first port of call for knowledge is the Google search engine. In a recent report on information seeking patterns, commissioned by the British Library and JISC (Joint Information Systems Committee), the authors explore the myths around the Google generation and point out that "in a real sense, we are all Google generation now" (Rowlands et al. 2008: 301). There now exists convincing evidence from libraries' deep log statistics that people of all ages use the Internet and its various technologies in surprisingly simple ways. Indeed, the investigation shows that the digital information world is characterized by massive choice, easy access, and simple to use tools, but people "from undergraduates to professors" exhibit "a strong tendency towards shallow, horizontal, 'flicking' behavior in digital libraries" (Rowlands et al. 2008: 300).

The second reason stems from a debate on quantities of empirical evidence in English historical linguistics. Some of the articles in a recent handbook (Nevalainen & Traugott, eds. 2012) deal with observing recent grammatical change and focus on the benefits and disadvantages of corpus materials of varying sizes. Davies (2012), a proponent of large corpora, illustrates that the value of large (but highly structured) corpora such as the *Corpus of Historical American English* (COHA) lies in the fact that they enable looking at not only low frequency lexical and semantic changes, but also low/medium frequency lexico-grammatical structures that may indeed reveal the entire life cycle of a change from its incipient stages onwards. He argues that the advantages of a large historical corpus, such as COHA, come from the fact that it is a balanced collection in which the conversion of the materials has been carried out accurately. In addition, it contains an enormous number (100,000+) of texts annotated for year of publication and genre, and all this comes equipped with a web-based search interface that enables a range of searches.

As for the advocates of small and structured corpora, such as the Brown family of corpora or diachronic materials such as ARCHER, Hundt & Leech (2012) show that the particular strengths of smaller materials have to do with careful sampling, exhaustive, qualitative scrutiny of the raw results, whole-text access and accurate metalinguistic information of the materials. They also point out that equidistant observation points of 30 years, roughly equalling to one generation, make these corpora suitable for observing language change. They provide two case studies both of which require careful qualitative scrutiny of the raw results for excluding homonymous forms (the first of relative pronouns in restrictive relative clauses and the second one of *for* as a causal conjunction).

It goes without saying that Google Books and its Ngram Viewer do not represent a structured historical corpus *per se* even though it can be "an incredibly useful tool for looking at the frequency of exact words and phrases" as Davies (2012: 159) points out. It is an unorganized text archive that offers access to massive amounts of language material in a nice, easy-to-use package. This source has been used in a range of studies. They include for instance investigations in variationist sociolinguistics (Tagliamonte 2016), historical pragmatics (Jucker et al. 2012), English historical linguistics (Friginal et al. 2014), and psycholinguistics (Brysbaert, Keuleers & New 2011). In addition, Google Books is recommended as a useful source for trend studies across time in course books in sociolinguistics (cf. Friginal & Hardy 2014: 188).

Since Google Books is one of the closest sources of evidence to hundred-million word mega corpora (such as COHA) and offers easy and open access to hundreds of millions of words of texts of British English (and other English varieties and other languages) that correspond with the time frame of the CEECE, it is clearly justified to ask what kinds of evidence can be found in it by someone interested in language history and language change in social-historical context. And most importantly,

how does this information fare with the evidence from carefully sampled corpus resources, such as the CEECE? Our discussion deals with Google Books, but there are other text mining tools, such as the EEBO N-gram Browser (available at http://earlyprint.wustl.edu/), which cover English print culture up to 1700.

The usefulness of Google Books as a source for data is discussed in Nevalainen (2013) who studies the diachronic evolution of three sets of words related to polite society. These "buzz words" (*courteous/courtesy, civil/civility, polite/politeness*) have distinct histories in the Middle Ages, the Renaissance, and the Enlightenment, and Nevalainen looks into the evidence provided by Google Books' Ngram Viewer. She points out that one obvious shortcoming in this vast database is the fact that the materials are unevenly distributed and the early periods are considerably smaller than later ones. She concludes that at the time of multiple data sources, i.e. small structured and large corpora, and big/messy databases, linguistic study reaches the best result by triangulating a range of evidence from data sources.

Lijffijt, Säily & Nevalainen (2012) explore culturally loaded words over time, and focus on the frequencies of war-related vocabulary during the seventeenth century. Their results from Google Books show, unexpectedly, that the frequency of the lexeme *war* peaks during the Civil War (1642–1651), but that its frequency continues to grow even after the war toward the end of the century. They conclude that big and messy databases can be useful heuristic tools in diachronic studies, but that the results obtained should be checked against more reliable data using significance testing. They also point out that improving the reliability of the Google Books data requires analysing the actual n-gram frequencies aggregated over longer time periods.

Pechenick et al. (2015) and Koplenig (2017) draw attention to the scarcity and dubious quality of the metadata associated with Google Books. For instance, Pechenik et al. (2015) show that the 2009 version of the dataset labelled "English Fiction" in fact seems to consist mainly of scientific writing. While the 2012 version represents fiction more accurately, it is clear that the Google Books metadata is still far from perfect. Crucially for diachronic studies, we have observed that the year of publication is often inferred incorrectly, so that a book published in the twentieth century whose topic concerns the sixteenth century may be falsely categorized as a sixteenth-century publication. Furthermore, both Pechenick et al. (2015) and Koplenig (2017) note that as the composition of the corpus and how it changes over time is largely unknown, it is difficult to discover whether a change in frequency represents actual linguistic or cultural change, or whether it is merely an artefact of the data. Again, triangulation with more reliable sources is called for (Nevalainen 2013).

From these previous studies concentrating primarily on lexical variation, this section shifts the focus towards the focus in the book, *viz*. lexico-grammatical

variability in social context. One way to illustrate our approach is to explore what types of evidence can be extracted from Google Books using a variable which is discussed more extensively in one of the chapters. The variable concerns the various ways of expressing singular human reference using indefinite pronouns. In Early Modern English, the variable consisted of four variant forms, the older -MAN and the independent forms which were decreasing in frequency, and the incoming -ONE and -BODY indefinites, as has been shown by Raumolin-Brunberg & Kahlas-Tarkka (1997). During the 15th–17th centuries, before spelling in general became standardized, the spelling for the determiners (EVERY, ANY, SOME, NO) varied considerably. As an illustration from the CEEC, it is easy to find nearly two dozen spelling forms of a determiner form ANY which can be used in compound indefinites (e.g. *any, ani, anie, anii, anney, anny, annye, anye, ene, eney, enu, enni, anny, eny, ony, onny, onnye, onye*). In addition, spelling variation exists in highly literate texts, as shown in (13.1) from the early 17th century:

(13.1) That was the true light, which lighted **euery man** that cometh into the world
(The *Authorized Version* of the Bible, John 1:9)

As for the long eighteenth century, the orthographic variation is mainly related to spelling the indefinites as separate units. According to Denison (1998: 101), it is only in the early nineteenth century when the indefinites start appearing as one unit. This means that for a comprehensive overview of the indefinite pronoun use from Early Modern English to the present day, one needs to account for the spelling variants in both joined and separate forms.

The searches in Google Books offer limited tools for accounting for the spelling variation, particularly since many of the spelling variants are extremely low frequency items. The only feasible tool in the search interface is the sum operator "+" that allows combining multiple n-gram series into one, but it cannot be used in case insensitive searches so any query must contain both lower and upper case forms. Despite the limitation, the operator function enables combining the searches for the most common spelling variants (i.e. for the assertive paradigm: ANY ONE + ANYONE, ANY MAN + ANYMAN, etc.) and this search needs to be repeated for all four paradigms.

Figure 13.3 illustrates the pooled results depicting the development of the three nominal forms in the indefinite pronouns from the early 17th century onwards. It shows the results of the relative frequency for the compound indefinites with -ONE, -BODY and -MAN endings, but not for the independent forms that are homonymous with pro forms in the negative paradigm (*He asked for money, but I gave him none*).

The diachronic picture that emerges from the search in this large unstructured database is twofold. Firstly, the decrease of the forms in -MAN, which had been decreasing in frequency from the 15th century, is also visible in the Google Books

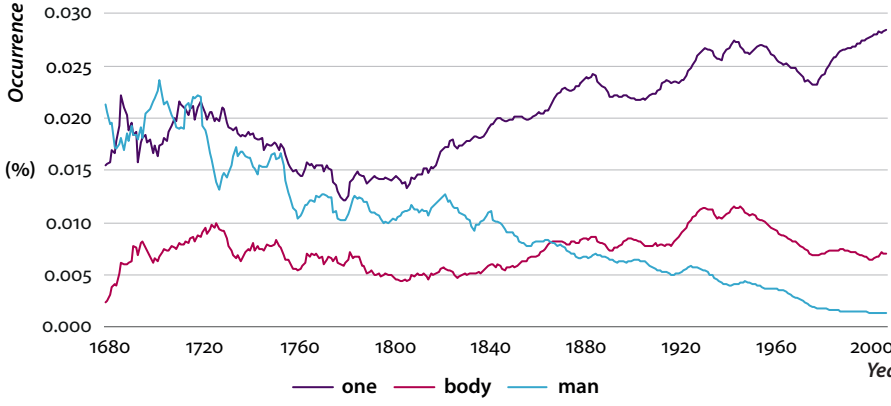

Figure 13.3 -ONE, -BODY and -MAN indefinites in British English Google Ngram Viewer

results. The cross-over from -MAN to the incoming variant forms takes place in the early 18th century. This result is partly similar with the results in Nevalainen & Raumolin-Brunberg (2003: Table 14 in Appendix II) in which the crossover in correspondence data takes place in the mid-17th century. Secondly, what is noteworthy is that the crossover takes place to the -ONE forms rather than to -BODY. Previous studies, which have drawn evidence from smaller corpora which make it possible to circumscribe the variable and enable careful qualitative investigation of each occurrence, have shown that it is the -BODY variant which emerges earlier than -ONE (Raumolin-Brunberg & Kahlas-Tarkka 1997: 68 using the multi-genre *Helsinki Corpus of English Texts*). This observation is most likely brought about by the fact that the Google Books results also include partitive OF-forms in which only -ONE can be used but not -BODY (cf. ANYONE OF THEM, but *ANYBODY OF THEM). Similarly, the results also include both generic and non-generic cases.

It is obvious that even this diachronic overview of a very simple variable of compound indefinite pronouns from the Google n-grams requires additional corpus evidence, and the results obtained are only partially accurate provided that one knows the limitations in the searches. These limitations relate to excluding (a) partitive structures which skew the picture, and (b) the independent forms that would need to be found through extensive post-editing of the raw results which is not possible in Google Books. So Davies's (2012) conclusion, that Google's resources are inadequate for diachronic linguistics since they only enable looking at the frequencies of exact words and phrases, is fully justified, but it should be added that the results can be used as a confirmatory tool supplementing more fine-grained data.

In the case of the indefinite pronouns with human generic reference, this fine-tuning is related to two issues. The first one comes from Hundt & Leech's (2012) notion of exhaustive scrutiny for qualitative analyses and is clearly shown

in the results above. They include a considerable share of false positives, and their precision is thus limited even though the results are based on a large set of data. The second one is related to the sociolinguistic embedding of the variant forms. Sociolinguistic variation in the indefinite pronoun use is important because the variable consists of two incoming forms (-ONE and -BODY) and two outgoing ones, one of which is semantically loaded for gender-specific references (i.e. -MAN).

In previous studies (Nevalainen & Raumolin-Brunberg 2003), the early modern period variability was found to correlate with writers' gender above all the other social categories. Zooming into sociolinguistic variability and comparing the findings with a range of other structures in Part II of this book has offered answers to whether this development continues in the later stages of the change (cf. Chapter 15 below). In addition, information of sociolinguistic embedding could help to answer the question of how linguistic variables may index social characteristics and reveal how speakers place themselves within the social landscape through their linguistic practices (cf. the third wave of sociolinguistics in Eckert 2012). For instance, information on the frequencies of outgoing variants may enable identifying characteristics such as linguistic conservatism, i.e. speakers who lag behind the others in linguistic change and who polarize the process (see Chapter 14 below). It goes without saying that the evidence from a large unstructured database in this case falls short of such a close analysis.

Another limitation in the big and messy database of Google Books concerns the nature of the data included and the question of its representativeness. While Google Books contains a massive number of books, not all periods are equally well represented in the database. For instance, the British English section in the 2012 version of the database[1] goes back as far as the 16th century, but the amount of data in that century is extremely limited and patchy, so that even the frequency of the most common lexical items in the language, such as *the*, cannot be studied reliably using the Ngram Viewer. This is partly due to the fact that the viewer employs per-year frequencies and lacks the basic feature of pooling the data into longer time periods. As there is data from only 38 years in the 16th century, for the remaining 61 years the frequency of *the*, as well as that of any other item, is zero. If smoothing is applied to the graph (the default smoothing of 3 means that the yearly frequency is averaged across three consecutive years), the smoothed frequency will be erroneously low, because the zeros are not treated as missing data points but as actual frequencies (see Lijffijt et al. 2012: Section 2.2).

The situation with the 17th-century data is better in that every year except for 1601 is represented by at least one book, making it possible to study high-frequency

1. Statistics on various versions of the database, as well as the datasets themselves, can be downloaded at http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

items. This is an improvement over the 2009 version of the database, where Lijffijt et al. (2012: Section 2.2) report 15 empty years between 1600 and 1640. From 1680 onwards, which is the starting point of this volume, each year is represented by at least fifteen books. In the 1700s, the amount of data ranges from 64 to 731 books per year for a total of 29,034 books, but rare or poorly dispersed items remain difficult to study using the yearly approach. It is only in the 19th century that the truly big data begins to appear, the 19th–21st centuries being all represented by thousands of books per year for a total of hundreds of thousands of books per century (see Table 13.1).

Table 13.1 Number of volumes by century in the British English section of the Google Books database (2012 version). Based on yearly totals retrieved from http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

Period	Number of books
1500–1599	141
1600–1699	1,645
1700–1799	29,034
1800–1899	286,693
1900–1999	563,980
2000–2009	232,384
Total	1,113,877

The basic Ngram Viewer search enables the user to search for one or more words or phrases (i.e. n-grams), after which the results are displayed as a line graph with time on the *x*-axis and relative frequency on the *y*-axis (Google 2013b). As noted above, words (such as different spellings of the same word) can be combined into a single n-gram series by using the + operator. Phrases may include one instance of the wildcard *, which stands for any word; the graph will then show results for the top ten words, which is perhaps not what a linguist would want. Furthermore, the * operator cannot be used word-internally to replace a sequence of characters, which means that the viewer cannot be used to search for affixes. Part-of-speech tags can be used at the end of words (e.g. cook_VERB) or on their own (_VERB_ or *_VERB), but their use is severely restricted as they cannot be utilized in queries that contain a wildcard standing on its own or in queries consisting of more than three words. Similar restrictions apply to the lemma query operator _INF (cook_INF comprises the inflected forms *cook*, *cooks*, *cooking*, *cooked*) and to the dependency relations operator => (dessert => tasty will reveal how often *tasty* modifies *dessert*). As pointed out above, case-independent search is only available for simple queries that do not contain wildcards for instance.

To address some of these issues, Mark Davies (2011–, 2014) has developed an advanced interface for the Google Books data. However, the amount of pre-19th century British English data in Davies's version appears to be considerably smaller than in the official interface, and it is unclear what has been excluded. In Davies's version, too, there are limitations on what kinds of query operators can be combined. Moreover, the part-of-speech annotation in Davies's version is based on matching individual words with their most frequent parts of speech in a corpus of present-day American English, which means that it is less reliable than the official annotation, which is based on analysis of the running text of the books (Lin et al. 2012). Because of these limitations, the following analysis will employ the official interface.

Let us now turn to the rest of the changes analysed in this book. How well can they be studied using the Ngram Viewer and its British English component? First, it is clear that the interface is of no use to the study of suffixal productivity in Chapter 12 as searching for affixes is impossible. This also means that the verbal *-s* (Chapter 7) and progressive *-ing* (Chapter 11) can only be studied by searching for some of the most frequent verbs in which they occur, which is obviously inadequate but may reveal something of interest. Our findings are described further in Figures 13.4–13.8.

The Ngram Viewer interface does not enable disambiguating singular and plural *you*, and Figure 13.4 therefore focuses on forms of *thou* alone. The figure shows the overall decline in the use of *thou* towards the present. As expected, the number of instances is very low from the 1680s onwards. The occasional peaks can be explained when we look more closely into the material used by the Ngram Viewer: most of the data come from different prints of the Bible and the works of Shakespeare, both of which traditionally make a distinction between *thou* and *you*.

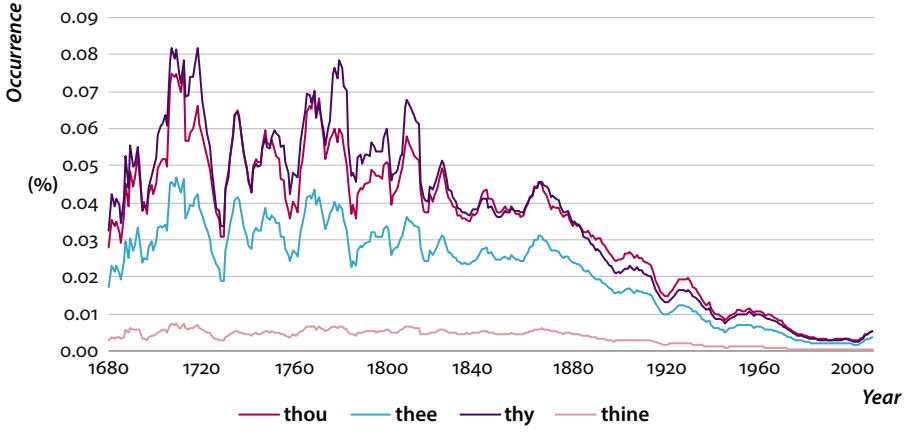

Figure 13.4 Forms of *thou*

Since the interface does not allow searching for affixes, Figure 13.5 focuses on a single frequent item, *has*, vs. the older form, *hath*. The Google Books figures agree with both the EEBO-TCP and CEECE data discussed in Chapter 7 that the crossing-over from *has* to *hath* took place at the very end of the 17th century. However, as expected, the mixed print materials included in the Google Books suggest a much longer tail for the use of the outgoing *hath* variant in the post-1700s than could be seen in personal correspondence.

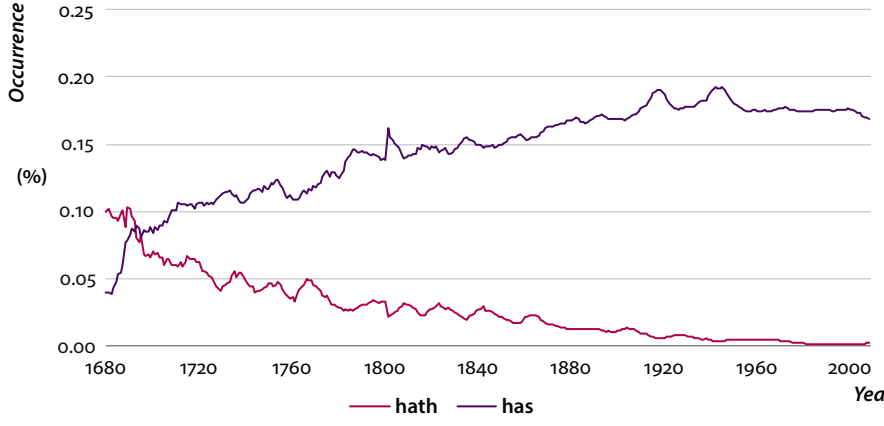

Figure 13.5 Diffusion of verbal *-s*

As the general case cannot be reliably identified using the Ngram Viewer, Figure 13.6 focuses on *do* followed by five of its most frequent lexical verb companions in

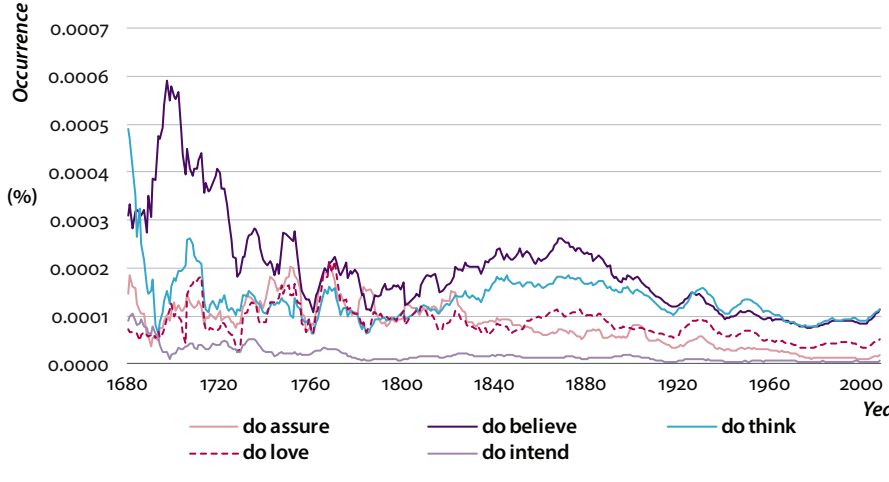

Figure 13.6 Periphrastic DO in affirmative statements

the CEECE. The results from the Ngram Viewer support the general view of the development of periphrastic DO in affirmative statements: there is a clearly declining trend from the early eighteenth century onwards. Nevertheless, the construction never goes completely out of style. Biber et al. (1999: 433–434) list several high-frequency verbs that still go with DO today (see Section 8.2.3 above). Of the ones tested here, only *think* and *believe* are included in their list, which agrees with the findings in Figure 13.6, as the other three verbs are clearly less frequent by 2008. It can also be argued that *assure* may have been genre-specific even in the eighteenth century; Biber et al. certainly identify register variation in verbs appearing with DO. It might be an interesting exercise to trace the development of DO with all the verbs listed by Biber et al., but this would undoubtedly leave aside interesting developments with other verbs, genre-specific trends and, of course, cases where there is an intervening adverbial (see Sections 8.2.4 and 8.4.3 above). For cases like periphrastic DO in affirmative statements, the Ngram Viewer is more a source of supporting information than even a diagnostic tool.

Since the Ngram Viewer makes it possible to use part-of-speech tags, the search on the third-person neuter possessives in Google Books (Figure 13.7) focuses almost on the same grammatical contexts as Palander-Collin (see Section 10.2 above). The constructions *its N*, *the N of it* and *the N thereof* can be compared, although *the N of it* has to be shortened to *N of it* owing to the limitation of part-of-speech tags to queries of a maximum of three words. The overall trends are highly similar with the findings based on CEECE. In the CEECE, *its N* rises during the eighteenth century and reaches 76% of the paradigm towards the end of the period studied (1780–1800; see Section 10.4.1). Figure 13.7 shows that the frequency of *its N* further increases, as would be expected, and peaks around 1840s.

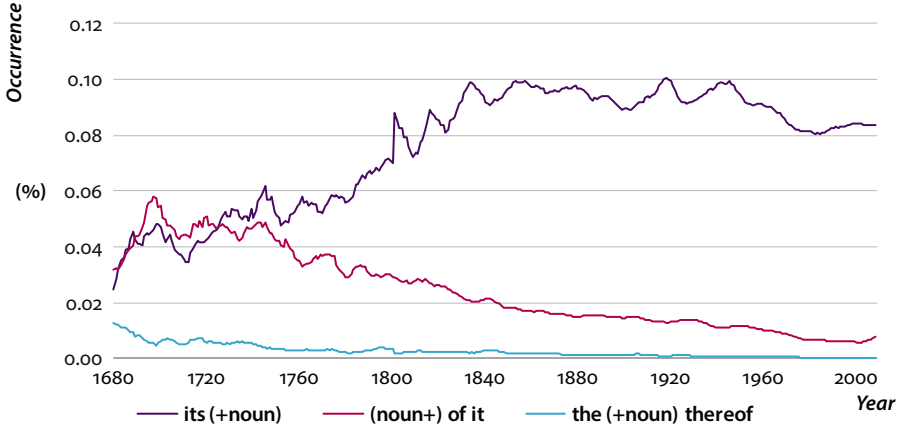

Figure 13.7 Diffusion of *its*

Because the Ngram Viewer interface does not allow affix searches, the changes in the progressive aspect are illustrated in Figure 13.8 through a verb construction that is frequent in letters, namely BE *writing/Writing* (*I am writing*). The numbers are very low, but the increase of this particular verb form follows the observed upward trajectory of the progressive from Early Modern English to the present day (see Kranich 2010: 95 for an overview).

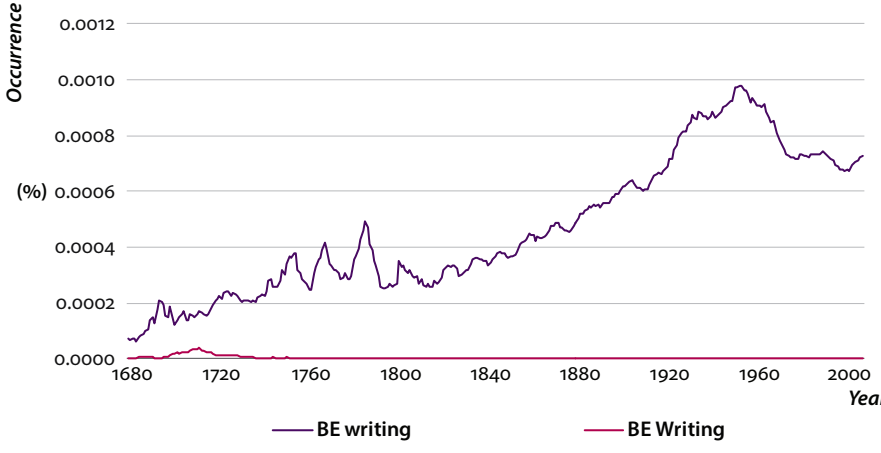

Figure 13.8 The progressive aspect: BE *writing* (lower- and uppercase spelling)

To conclude, we have seen that Google Books may provide a useful point of comparison or starting point for the study of some grammatical changes. However, the limitations of the interface – and the messy, patchy and uncertain nature of the data and the metadata behind it – mean that Google Books alone is not a sufficient source of data. This is especially true in the case of sociolinguistic research, where we need to know whose language it is that we are analysing, which groups lead the change and which are lagging behind. For this we need carefully-compiled specialist corpora with rich sociolinguistic metadata, such as the CEECE.