Language prescriptivism : attitudes to usage vs. actual language use in American English

Kostadinova, V.

CHAPTER 6

Patterns in actual language use

6.1 Introduction

In Chapter 4, I explained the general approach taken in this study to exploring the question of whether prescriptive metalinguistic discourse affects usage patterns both across time and across register. I also explained that this will be approached by comparing patterns of change observed in the treatment in usage guides of the six linguistic features investigated, i.e. precept (see Section 4.2), with patterns of variation in the actual use of those linguistic features, i.e. practice. Having analysed the precept data in Chapter 5, I now turn to the patterns of actual use of each of the six features, i.e. *ain't*, the discourse particle *like*, *literally*, negative concord, pronouns in coordinated phrases (i.e. object *I* and subject *me*), and the split infinitive.

The data on actual use are taken from the two large-scale corpora introduced in Section 4.4, COCA and COHA. In that section, I also explained that the patterns of language use will be explored on the basis of two analytical approaches, or two types of metrics (cf. Biber et al. 2016). First, I look at the patterns of variation by identifying the text-linguistic frequency of occurrence of linguistic variants considered problematic to varying degrees from a prescriptive point of view, i.e. *ain't*, the discourse particle *like*, the non-literal use of *literally*, negative concord, object *I* and subject *me*, and split infinitives. Secondly, I use the variationist approach to analyse the proportion of the use of some of these variants in the context of their linguistic variables by identifying the proportion of use of the unacceptable linguistic variant out of the total number of environments in which it could occur. For example, I look at the proportion of *ain't* for *be not* out of the total number of environments in which a *be not* variant is used, or the proportion of split infinitives out of the total number of infinitives modified by a single adverb, both split and not split. For more details on the identification, extraction, and disambiguation of the occurrences for each of the features, see Section 4.4 and Appendix C. Sections 6.2 – 6.7 discuss the patterns of occurrence of each of the six linguistic features across time periods and the various corpus genres: academic, fiction, magazines, newspapers, and spoken (see also Table 4.2).

In addition to this, I present an analysis which aims to empirically identify the potential influence of prescriptivism on the use of the split infinitive. Using this feature as a case study, I conduct a multifactorial analysis, in order to identify the extent to which the use of split infinitives is associated with the use of other prescriptively targeted features, at the level of individual texts. Section 6.8 presents the results of this analysis. In the final section, I bring these findings together, and discuss the issue of the influence of prescriptivism on language use.

6.2 *Ain't*

did, or with other auxiliaries and modals.

As explained in the previous section, for the purposes of this analysis I rely on two types of metrics in order to analyse the patterns of usage of *ain't* across time periods and genres in the corpora. The first account of the patterns of use of *ain't* is the normalised frequency of use of all occurrences of *ain't* in the corpus, irrespective of their function. The reason that this may be considered a good indicator of the changing patterns of usage of *ain't* is that, regardless of the *function* of *ain't*, the *form* is generally stigmatised. The second metric measures the proportion of *ain't* used for *be not*, in the context of all possible environments of *be not*, as well as *ain't* used for *have not*, in the context of all possible environments of *have not*. The reason for the second type of metric is that, despite the general stigmatisation of *ain't*, there is a sense of *ain't* for *be not* being somewhat more acceptable than *ain't* for *have not*. In order to explore the extent to which such ideas identified in the precept data relate to patterns of actual language use. The variables used in the analysis are given in Table 4.3.

In Section 3.3, I briefly outlined the major findings from previous studies on the variation in the use of *ain't* in American English, in terms of linguistic and sociolinguistic constraints. The complex variation in the use of *ain't* is reflected in the data analysed for this study. First, with reference to the linguistic variation in the use of *ain't*, the analysis showed that alongside the predominant uses of *ain't* in environments of *be not*, in examples (48) and (49), and *have not*, as in (50), there were a number of cases where *ain't* is used as a variant of *didn't*, as in (51), with modals such as *mustn't*, as in (52), and possibly, in a small number of cases, with *wasn't* (cf. Anderwald 2002). Finally, what is an interesting and, I believe, significant finding resulting from the corpus data was the discovery of a number of occurrences of a metalinguistic mention of *ain't*, in (53), in which the word is criticised or implicitly associated with the proscription against its use. These will be discussed in more detail in the final part of this section.

(48) He thinks he **ain't** a man any more. (1987, fiction, COHA)

(49) He **ain't** saying that to my face. (2006, spoken, COCA)

(50) You **ain't** said yes yet. (1932, fiction, COHA)

(51) Why y'all **ain't** call me? (2011, magazine, COCA)

(52) You must be joking, **ain't** you, Mr Luther? (1940, fiction, COHA)

(53) Language of this sort could be terrifying to someone who only the week before at Miss Burke's had been sent to detention for saying **ain't**. (1959, fiction, COHA)

The data thus confirm previous accounts of the variation in the uses of *ain't*; however, it also confirms that most of these uses, exemplified in (48)–(53), are fairly rare, even in non-standard spoken data. Since the corpus data used for the present analysis reflect the standard American language variety, it is not surprising that these variants are very rare. This means that, despite the existence of the different variants, the greatest majority of *ain't* uses are found in the environments for *be not* and, to a lesser extent, *have not*. As a result, all other cases were excluded from the variationist analysis presented here.

The normalised frequency distributions of all occurrences of *ain't* across time periods are shown in Figure 6.1. The figure contains two subfigures, one for the

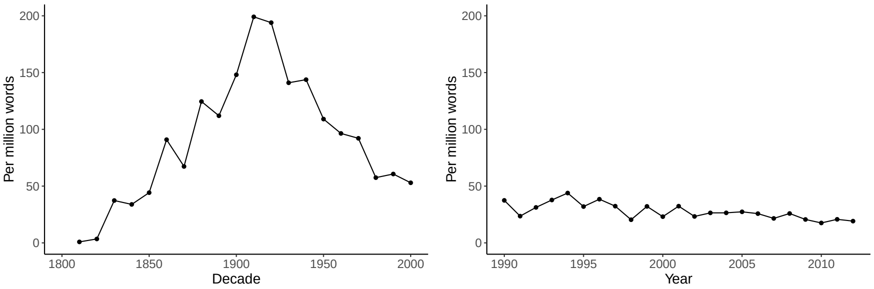

Figure 6.1: Text-linguistic frequencies of all occurrences of *ain't* across time (COHA: n = 39,348; COCA: n = 12,228)

rate of occurrence of *ain't* in COHA, and the other for the rate of occurrence of *ain't* in COCA.[1] Due to the make-up of the corpora, as well as the proportionally different time scales they cover, the time periods used for COHA are decades, while those for COCA are years. The second subfigure can thus be seen as zooming in on the last two decades in the period under investigation. As the graphs show, the frequency distribution of *ain't* undergoes a striking increase until the 1910s, followed by a similarly dramatic decline in the course of the twentieth century. Since the year 2000, the frequency of *ain't* has remained steadily low. While these results might lead us to postulate that prescriptivism may have had some effect on the use of *ain't*, it is important to consider other factors first.

One of those factors is register variation, which I also explore using both text-linguistic and variationist metrics to establish the normalised frequencies and proportions of *ain't* across the subsections of the two corpora used. The results from the text-linguistic analysis are given in Figure 6.2, which shows the normalised frequencies of occurrence of all cases of *ain't* across sections of the two corpora. The vertical axis represents the number of occurrences of *ain't* per million words across the major genre sections of the two corpora, i.e. fiction, magazine, newspaper, and non-fiction in COHA, and academic, fiction, magazine, newspaper, and spoken in COCA, which are plotted on the horizontal axis. The two plots show that the frequency of occurrence of *ain't* is highest in fiction in both corpora, with the fiction section in COHA containing the highest rate of occurrence of *ain't*.

[1] As evident from the graphs, the two corpora overlap for the period 1990–2000. There is some overlap in the materials included in the two corpora for the final decade of the twentieth century. For transparency, I represent the figures in their entirety, as well as separately, due to the fact that the make-up of the corpora is not entirely the same.

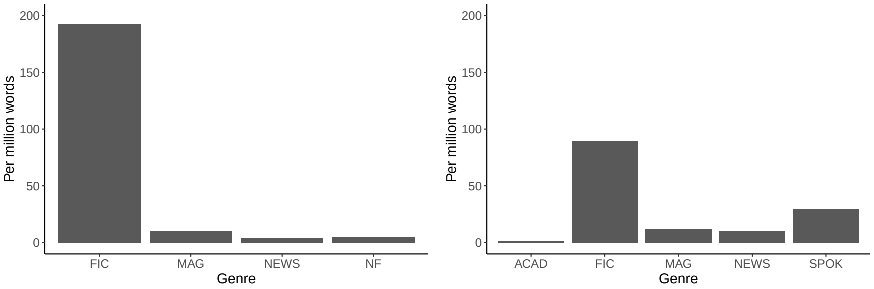

Figure 6.2: Text-linguistic frequencies of all occurrences of *ain't* across corpus sections (COHA: n = 39,348; COCA: n = 12,228)

Since the results of the effects of genre on the use of *ain't* show that the form is especially frequently found in fiction, I also plotted the trends for the occurrence of *ain't* in all other corpus genres taken together, excluding fiction. These results are given in Figure 6.3. There is a clearly even trend, with almost no difference whatsoever in the normalised frequency of occurrence over the course of the entire period investigated. There is a very slight increase at the end of the twentieth century, which could perhaps partly be explained by the presence of spoken data in COCA. A comparison between Figures 6.1 and 6.3 confirms the fact that the large-scale increase observed over time in the frequency of occurrence of *ain't* in COHA is an effect of its increase in fiction.

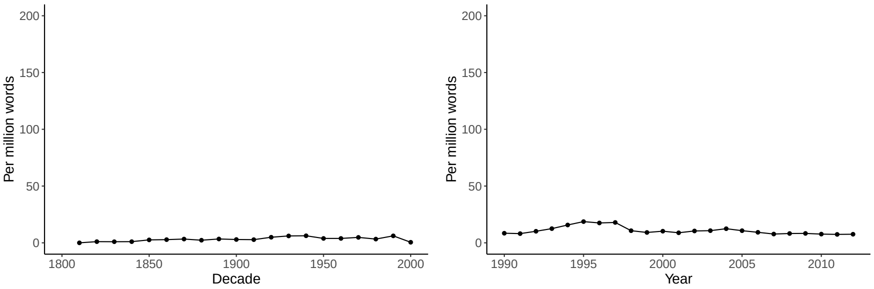

Figure 6.3: Text-linguistic frequencies of all occurrences of *ain't* across time, excluding fiction (COHA: n = 1,373; COCA: n = 4,751)

The question then is whether this increase and subsequent decrease in the rate of occurrence in fiction is a change in the use of *ain't* in this particular genre, or whether there are other explanations for the trend observed in Figure 6.1, such as the make-up of the fiction section. The latter scenario was investigated with further exploration of the make-up of COHA, focusing specifically on the subgenres included in the fiction section (see Table 4.2). This analysis shows that the percentage of drama texts out of all fiction texts is the highest for the 1910s (i.e. 33.90%) and the 1920s (i.e. 33.10%; see Table C.1 in Appendix C for the percentage of drama texts for the other decades in the corpus, which is lower than for the 1910s and the 1920s). Similarly, almost 30% of all occurrences of *ain't* in fiction in those two decades come from drama texts. This means that *ain't* is a feature which is characteristic of fiction in general, and of plays in particular. This in turn also suggests that the increase and decrease in the rate of occurrence of *ain't* in fiction is more likely to be related to the higher percentage of drama texts for those two decades, rather than being a consequence of changing patterns of usage.

Having established that there has been no change in the rate of occurrence of *ain't* in American English since the beginning of the nineteenth century, and that the variation patterns observed are the effect of register, I now turn to the question of how this finding relates to the change in treatment of *ain't*. I already pointed out in the discussion of the treatment of *ain't* in usage guides (see Section 5.3) that during the course of the twentieth century this feature was increasingly viewed as acceptable in restricted contexts. On the basis of these two analyses, it could of course be the case that there is no relationship between language use and usage guide treatment, and that the two developments identified here are independent of each other. However, given the salience of *ain't* both as a dialectal feature and as a usage problem, this seems unlikely. Rather, it seems more likely that usage guides have changed their treatment of *ain't* as a consequence of the low frequency of the form in general standard American English, as well as its stable place as a dialectal feature, mostly used in fiction, and especially drama. In order to explain how this relates to the usage guide treatment of *ain't*, it is important to look more closely at the kind of acceptability of *ain't* that is expressed in usage guides. We can observe that, while usage guide writers, especially in the second half of the twentieth century, tend to be more accepting of *ain't*, this acceptability is still restricted to a few contexts. These contexts include specific functions of *ain't* in marking non-standard or dialectal speech in works of fiction, the use of *ain't* in set phrases and idioms, and its use in popular songs. These functions, it seems, have become more stable over the course of time, resulting in the low overall frequency of *ain't*. It is precisely this kind of regularisation of the contexts of use of *ain't* that may have allowed for its higher acceptability in restricted contexts in usage guides. The use of *ain't* in drama may therefore be understood as the reason for the acceptance of *ain't* in restricted contexts. In other words, once the feature became very limited in frequency in general language use, and its use in fiction became stable, the need to proscribe *ain't* slowly disappeared. This also implies that there is a time lag in the change in treatment.

The discussion so far has been based only on the text-linguistic frequencies of occurrence of *ain't* across corpus sections. In order to gain a better understanding of the use of *ain't* in the context of the variables *be not* and *have not*, I turn to the variationist analysis of *ain't*, looking not only at how *ain't* is used in particular types of texts or periods of time, but also at how it is used in relation to the other variants for *be not* and *have not*. Figure 6.4 shows the proportion of cases realised with *ain't*, as opposed to all other cases of *be not*, realised by both full and contracted forms, across decades in COHA and years in COCA. Figure 6.5 shows similar proportions for *ain't* functioning as *have not*, as opposed to the total number of cases of *have not*.

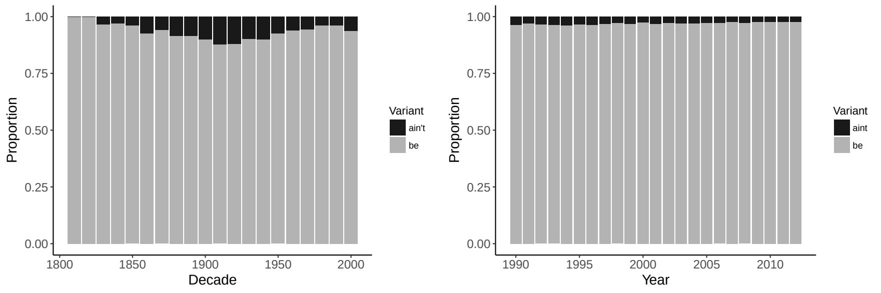

Figure 6.4: Proportion of occurrences of *ain't* (COHA: n = 30,106; COCA: n = 10,154) across time out of the total number of *be not* environments (COHA: n = 415,677; COCA: n = 637,133)

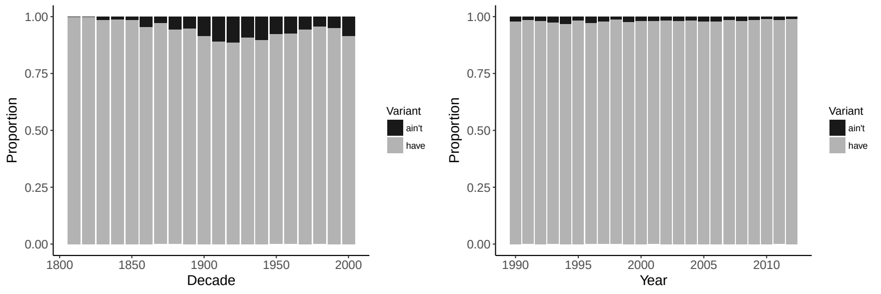

Figure 6.5: Proportion of occurrences of *ain't* (COHA: n = 6,762; COCA: n = 2,061) across time out of the total number of *have not* environments (COHA: n = 102,584; COCA: n = 110,890)

The variationist analysis of *ain't* for *be not* shows that the percentage of *ain't* was somewhat higher in the beginning of the twentieth century, as shown in Figure 6.4, reflecting the increase in frequency observed in the overall distribution of *ain't* in Figure 6.1. The figures for *ain't* for *have not*, given in Figure 6.5, are somewhat lower than *ain't* for *be not* in the historical data, and not much different in the contemporary data.

Turning to the patterns of variation across genre sections of the corpora, Figures 6.6 and 6.7 plot the proportions of *ain't* occurrences from the total number of possible environments in the context of the variables *be not* and *have not*, respectively. The distribution of uses of *ain't* for *be not* and *ain't* for *have not* across genres shows that fiction is the genre where almost all uses of *ain't* are found. The proportion of *ain't* is slightly higher in the COHA data, but both plots show that the overall proportion of *ain't* is fairly low.

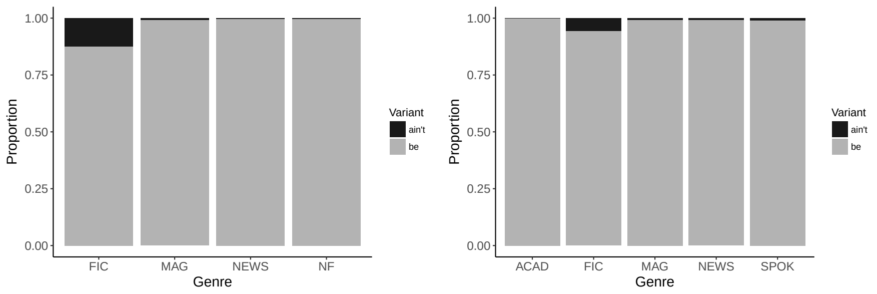

Figure 6.6: Proportion of occurrences of *ain't* (COHA: n = 30,106; COCA: n = 10,154) across corpus sections out of the total number of *be not* environments (COHA: n = 415,677; COCA: n = 637,133)

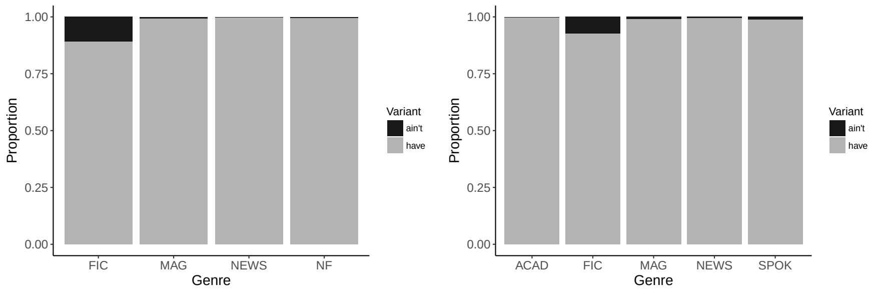

Figure 6.7: Proportion of occurrences of *ain't* (COHA: n = 6,762; COCA: n = 2,061) across corpus sections out of the total number of *be not* environments (COHA: n = 102,584; COCA: n = 110,890)

Having examined both the text-linguistic and the variationist frequencies of occurrence of *ain't*, I now turn to a brief discussion of the types of pronouns which *ain't* is most commonly used with. I explore this question in order to investigate whether there is any empirical basis for the high acceptability in usage guides of *ain't* with the first person singular, as opposed to its use with the third person singular. To illustrate this, I will focus only on the use of *ain't* for *be not* in COCA. In this dataset, in 38% of the cases *ain't* is used with something other than a personal pronoun, i.e. with an noun phrase headed by a noun or a proper noun. Of the remaining 62%, *I* is used in 17% of the cases, *it* in 19% of the cases, and *he* and *she* in 6% and 3% respectively. *I* thus appears to be only the second most frequent pronoun, after *it*. This evidence suggests that the prescriptive ideology concerning the acceptability of *ain't* is not supported by its actual use.

Finally, as mentioned at the beginning of this section, I found that in the corpus data used for the present study *ain't* also occurs in metalinguistic contexts, illustrated in examples (54)–(57) below. These examples testify to the stigmatised status of *ain't* and its association with non-standard speech. In some sense, then, these examples provide evidence at least for the cultural influence of prescriptivism, and certainly of the status of *ain't* as a usage problem.

(54) He looked up, clear-eyed to her pleasure, and wounded her with delight in the way he said, "I **ain't** done anything." "That's right," she said, nodding firmly. "Don't say **ain't**, just because I fergit now and then when I'm working hard, and haven't time for the fancies and the rights of this and that. But I don't want my baby-boy t'get habit of speaking wrongly." (1936, fiction, COHA)

(55) Language of this sort could be terrifying to someone who only the week before at Miss Burke's had been sent to detention for saying **ain't**. (1959, fiction, COHA)

(56) Or Lynn Smith Jr., a rancher who wears a cowboy hat, tucks pants into boots and still says '**ain't**'. (2000, newspaper, COCA)

(57) She wiped her eyes and gave Pelton a withering look. "Don't say '**ain't**'!" There is no such word… (1996, newspaper, COCA)

6.3 The discourse particle *like*

The frequencies of occurrence of the discourse particle *like* in COHA and COCA show a definite increase in the use of this feature over time. A variationist analysis of *like* was not attempted in this study, due to the difficulty of ascertaining the variable context in which the feature occurs and establishing the total number of potential environments (see Section 4.4). On the basis solely of text-linguistic frequencies, it can be noted that the discourse particle *like* has indeed seen a striking increase in occurrence in the corpus data analysed. This increase is particularly salient in the COCA data, which contain a spoken language section. Comparing this distribution to the usage guides' coverage and treatment of *like*, it is clear that the more likely phenomenon we are observing is that usage guide writers are reacting to a robust process of language change, which has also been accompanied by social stigmatisation, in some sense independent from the usage guide tradition.

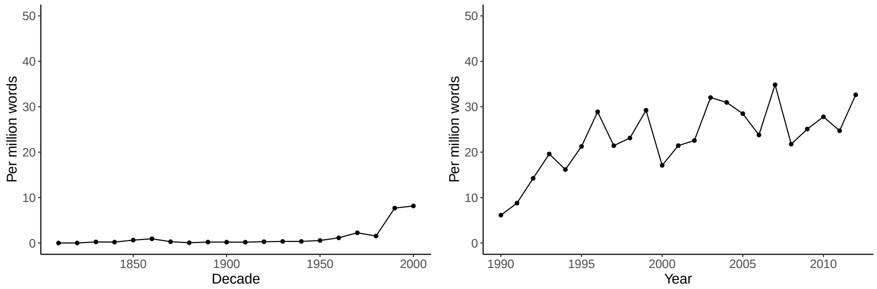

Figure 6.8: Text-linguistic frequency of the discourse particle *like* across time (COHA: n = 634; COCA: n = 10,020)

In terms of genre, Figure 6.9 shows that spoken data contain the highest number of instances of the discourse particle *like*, followed by fiction. Another important observation is the much higher frequencies observed for the COCA data, which reach almost 100 occurrences per million words, compared to 2.5 occurrences per million words for the highest frequency per genre observed in COHA. These distributions are hardly surprising, as the feature is a typical spoken language feature.

Both of these patterns of occurrence suggest that the case of the discourse particle *like* is a robust language change in progress, and is being led by the spoken language, as has been confirmed in many previous studies (see Section 3.4). This in turn provides further evidence that usage guides are responding to this development, which may suggest that *like* is on its way to becoming a usage problem. A crucial factor in this process, however, is the social stigmatisation of *like*, which preceded its treatment in usage guides. I return to these aspects of the use of the discourse particle *like* in Chapter 7 of this thesis.

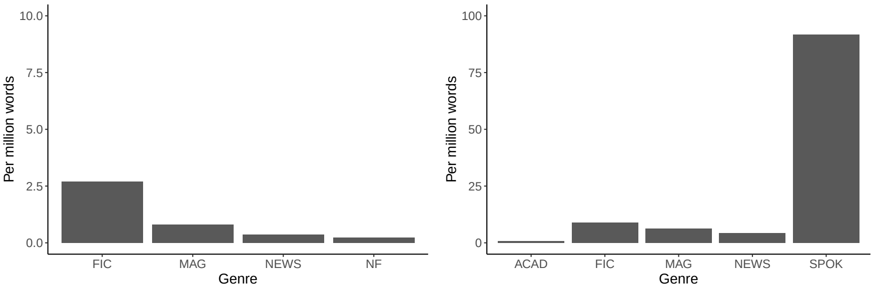

Figure 6.9: Text-linguistic frequencies of all occurrences of *like* across corpus sections (COHA: n = 634; COCA: n = 10,020)

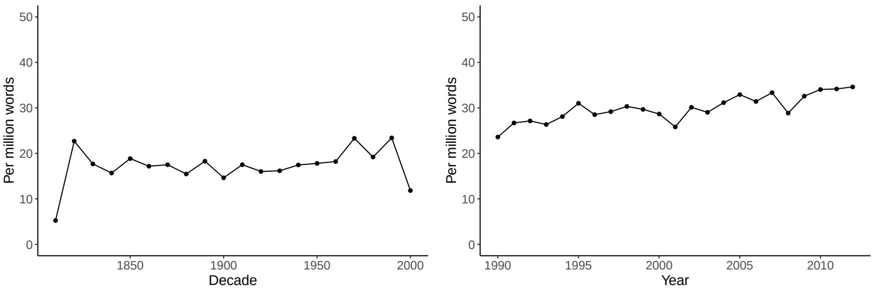

Figure 6.10: Text-linguistic frequency of *literally* across time (COHA: n = 6,848; COCA: n = 14,946)

6.4 Non-literal *literally*

The overall frequency of use of *literally*, as shown in Figure 6.10, has been increasing very slightly over the course of the last twenty years, from around 20 to a little more than 30 occurrences per million words, hardly a substantial increase. What these figures show, however, is that the notion that the word has come to be 'overused' is clearly not borne out by the data. In addition to plotting the overall frequency of occurrence of *literally*, two additional steps were taken in the analysis in order to arrive at a better understanding of the distribution of its three uses, as explained in Sections 3.5 and 4.4: primary use, dual use, and non-literal use.

As discussed in Section 4.4, the first step in the analysis was the automatic disambiguation of cases in which *literally* is used with its primary meaning from all other uses (see Appendix C for a description of the procedure). The results of this analysis are plotted in Figure 6.11, which shows the proportion of primary uses of *literally* as opposed to all other uses. The two graphs in the figure show that the use of *literally* in its primary meaning has remained fairly stable over time. The figure also

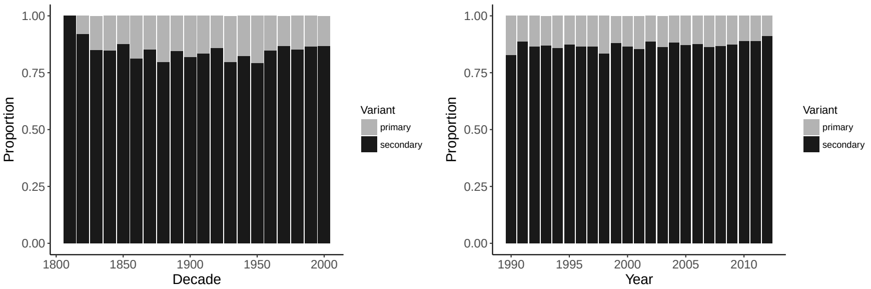

Figure 6.11: Proportion of primary uses of *literally* (COHA: n = 1,079; COCA: n = 1,937) across time compared to all other uses (COHA: n = 6,848; COCA: n = 14,946)

shows that the primary uses of *literally* are not the majority of the occurrences; rather, the reverse is the case. Note that this kind of automatic disambiguation, which is carried out using Python scripts, and relies on the part-of-speech tags in the corpus data, is bound to contain some degree of error in its precision and recall. In order to obtain a better picture of the rest of the uses of *literally*, as well as to supplement the automatic disambiguation, additional manual disambiguation was conducted on a sample of the total number of occurrences of *literally*, as described in Section 4.4. In this manual analysis I distinguished between the three uses of *literally*, *viz.* its primary, dual, and non-literal uses.

The results from the manual analysis are given in Figure 6.12, which plots the proportions of the three uses of *literally* across decades in COHA and years in COCA. A number of observations can be made on the basis of these trends. First, the graphs show that the number of primary uses of *literally* has decreased slightly over time. This is certainly the case if the distributions of *literally* in COHA and COCA are compared. It is worth comparing this figure with Figure 6.11, which shows the proportion of primary uses of *literally* against all other uses. The comparison shows that the difference between these two is in degree, but not in quality. This difference is not surprising, given that the automatic disambiguation is not as precise as manual analysis. However, it is reassuring that the patterns of distribution follow the same trend, which means that the automatic disambiguation is to a large extent reliable in tracking patterns of use. Secondly, it can also be observed that non-literal uses of *literally* are fairly rare, and that there has been little change in this respect in the course of the twentieth century.

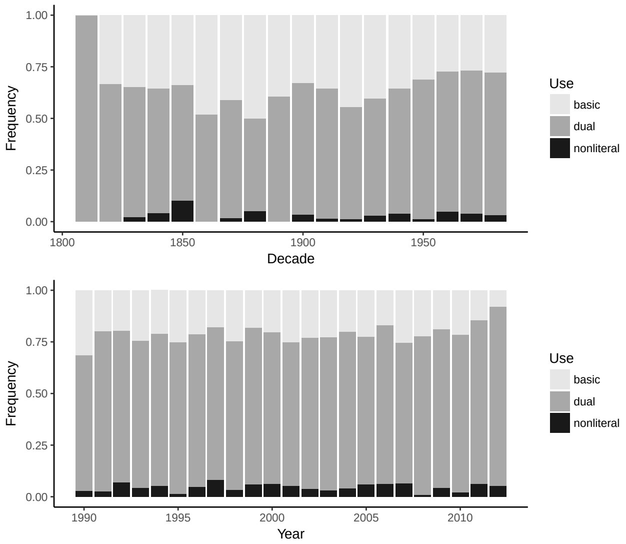

Figure 6.12: Proportion of the three different uses of *literally* across time, based on a sample of all occurrences of *literally* (COHA: n = 1,141; COCA: n = 2,864)

Finally, the dual uses of *literally* seem to be the most common, and the results of this analysis indicate that this was the case throughout the nineteenth and twentieth centuries. An interesting question is when in the history of the English language this use started to increase. I have not explored this question further, as the period before the nineteenth century is beyond the scope of this thesis, and previous studies on *literally* provide few corpus-based insights into its use in earlier periods. Some evidence on when the dual and non-literal uses of *literally* were first recorded can be found in the entry on *literally* in the *Oxford English Dictionary*;[2] on the basis of the instances recorded there, it can perhaps be hypothesised that these dual and non-literal uses of *literally* started to develop and to increase in frequency during the seventeenth century.

The frequency of occurrence of *literally* in all its uses across the genre sections in the corpora is given in Figure 6.13. The figure shows that the sections in COHA do not differ greatly in terms of frequency per million words. In COCA, the spoken section contains more occurrences of *literally* than any other sections. In both COCA and COHA, the fiction and the newspaper sections have the lowest frequency of occurrence of *literally*.

[2] See entry on *literally* in *OED Online*, available at www.oed.com.

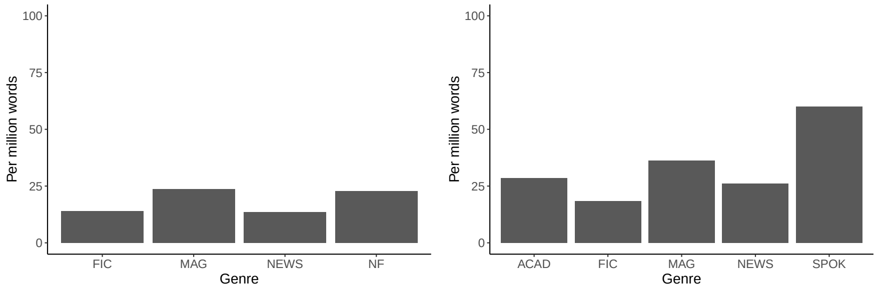

Figure 6.13: Text-linguistic frequencies of all occurrences of *literally* across corpus sections (COHA: n = 6,848; COCA: n = 14,946)

The analysis of the primary use of *literally*, as opposed to all other uses, shows that primary uses are highest in non-fiction texts in COHA and in academic texts in COCA (Figure 6.14). This distribution is expected, given that the primary meanings of *literally* are its oldest and the unproblematic uses. The manual disambiguation of a sample of these uses, the results of which are presented in Figure 6.15, shows a similar distribution pattern to that observed on the basis of the automatic disambiguation of the uses of *literally* plotted in Figure 6.14. A comparison between Figures 6.14 and 6.15 shows that the difference between these two is one of degree, rather than quality. For instance, for COHA, the non-fiction section has the highest proportion of primary uses of *literally*, followed by magazine, fiction, and newspaper; the differences are the same in both Figures 6.14 and 6.15, even though in Figure 6.14 the differences across corpus sections are less pronounced. This is likely the result of the fact that some relevant cases of the primary use of *literally* have not been identified using automatic disambiguation, based on part-of-speech tags. The difference between COHA and COCA which can be established on the basis of Figure 6.15 confirms that primary uses of *literally* are higher in frequency in COHA than in COCA, which might suggest a slow pace of change over time in the distribution of the uses of *literally*. Finally, Figure 6.15 also shows that non-literal uses of *literally* are very rare across all corpus sections.

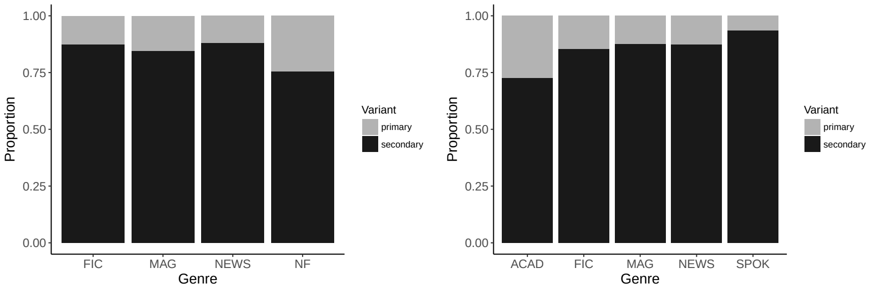

Figure 6.14: Proportion of primary uses of *literally* (COHA: n = 1,079; COCA: n = 1,937) across corpus sections out of all other uses (COHA: n = 6,848; COCA: n = 14,946)

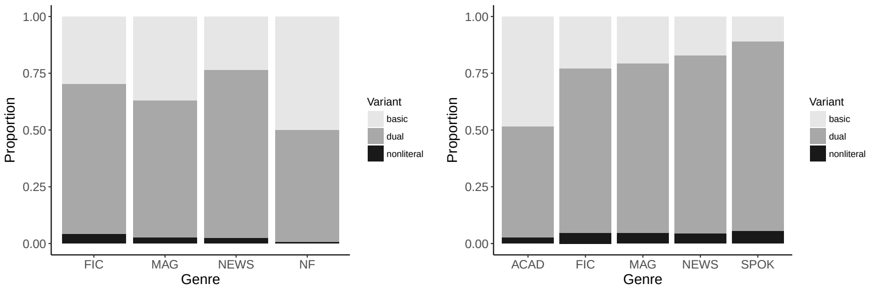

Figure 6.15: Proportion of uses of *literally* out of a sample of occurrences across corpus sections (COHA: n = 1,141; COCA: n = 2,864)

From these results it can be concluded that the frequency of the word *literally*, in all its uses, has not increased strikingly in the last 200 years, and that non-literal uses of *literally* are very rare. The primary uses of *literally* seem to have decreased somewhat in frequency in favour of its dual uses, although this change does not seem to be progressing rapidly. Comparing these results with the treatment of *literally* in usage guides leads to a number of observations. First, *literally* is a salient case of variation, and the extension of its meaning is considered problematic mostly due to the perceived opposition between its primary and its secondary uses (but see Powell 1992, who argues that there is a continuity of metalinguistic meaning underlying all uses of *literally*). Second, due to the salience of this process of variation and change, and perhaps in part due to the characteristic case of non-literal *literally*, this process has been interpreted by usage guide writers in a way which is not entirely supported by evidence from language use. The usage guide treatment of *literally* tends to distinguish between the 'strict' use of the word, which corresponds to its primary use, and all other uses, where *literally* is used either to express the opposite of its literal meaning, or simply to intensify an expression. However, the biggest problem with this kind of division is that the majority of the uses of *literally* are dual uses: these are cases in which it has both a literal meaning and an intensifying function. For example, in cases such as *There were literally millions of people*, the function of *literally* is both to express that there were more than one million people and to intensify the fact that this piece of information is surprising, and therefore worth emphasising. As a result of the lack of this kind of distinction, dual uses of *literally* may often be perceived as intensifying and superfluous, even though in principle they do satisfy the condition for the "proper" use of *literally*, in that in dual uses *literally* does not violate a literal reading.

In summary, what the case of *literally* shows is that salient cases of language variation and change may rise above the level of consciousness, and provoke metalinguistic discussions. It also shows that it takes time for a certain new language variant to rise above the level of consciousness before prescriptivists start noticing it (cf. Laitinen 2009). The same argument could be made for the case of the discourse particle *like*. Another aspect of the case of *literally* and the relationship between its status as a usage problem and its treatment in usage guides is that usage guide writers are in general mistaken in their overall characterisation of the use of *literally*. First, as I mentioned above, observations about an increase in frequency of the 'overuse' of the word are not supported by the data, which show a fairly stable and low increase in the frequency of use of *literally*. Second, the statements that *literally* has increasingly been used to mean precisely the opposite of its primary meaning are not supported by the data either: it is fairly clear that the incidence of non-literal uses of *literally* is very low, and has remained so for around two centuries. Finally, since *literally* is undergoing a slow process of change, which is at present perceived as an increase of variation in its meanings, what usage guide writers might be reacting to is the high number of dual uses of *literally*. In these uses, *literally* not only retains its literal meaning, but it also performs an intensifying function within an utterance. It may be these uses which contribute to the high salience of this feature, resulting in metalinguistic awareness and proscriptive commentary.

6.5 Negative concord

Negative concord is a non-standard vernacular feature, and this seems to be reflected in the very low frequency with which it is found in both COHA and COCA. The analysis of this feature was carried out on the basis of cases of negative concord with the three indefinites *no one*, *nobody*, and *nothing*. The use of Python scripts to identify and extract such occurrences in the corpora (see Appendix C for a description of the procedure) resulted in a dataset on the basis of which the frequency distribution of this feature is plotted in Figure 6.16. The figure shows that the normalised frequency of negative concord constructions with *no one*, *nobody*, and *nothing* is somewhat higher in COHA than in COCA; for COCA, the frequency has remained close to zero for the greater part of the last two and a half decades or so.

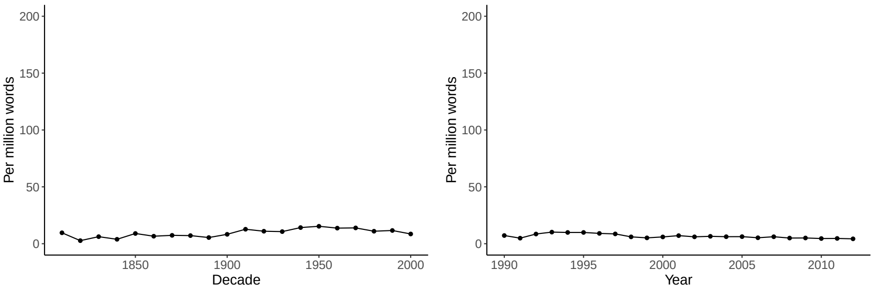

Figure 6.16: Text-linguistic frequency of negative concord across time (COHA: n = 3,912; COCA: n = 2,917)

The variationist analysis of negative concord identifies the proportion of negative concord with the three indefinites *no one*, *nobody*, and *nothing* of all potential uses of negative concord, by contrasting instances of negative concord with those of single negation with the indefinites *anyone*, *anybody*, and *anything* (cf. Nevalainen 2000). The results of this analysis are presented in Figure 6.17. The figure shows that the feature is not common in standard American English, with about 4% of all potential cases in both corpora being realised with negative concord. It is, however, worth noting that the three different indefinites exhibit slightly varying ratios of negative concord: cases of negative concord with *nobody* are found on average in 7.4% of all possible occurrences, compared to 4.2% for *nothing* and 2.4% for *no one*.

Turning to the examination of potential genre effects, Figure 6.18 shows the frequency per million words of negative concord across corpus sections. The

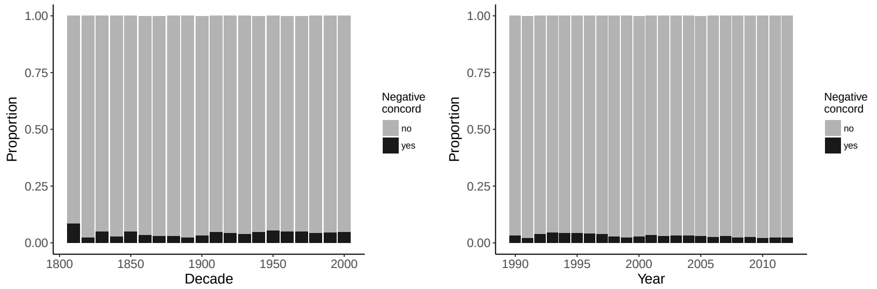

Figure 6.17: Proportion of occurrences of negative concord (COHA: n = 3,912; COCA: n = 2,917) across time out of the total number of environments for negation with the indeterminates *anything*, *anyone*, *anybody* (COHA: n = 91,165; COCA: n = 91,436)

frequency is indeed very low; while, like *ain't*, negative concord is limited to use in fiction in COHA, and fiction and spoken in COCA, the frequencies are lower than those of *ain't*. The feature is clearly not frequent in standard American English, and its uses are non-standard and limited to particular genres which are stylistically varied enough to contain higher levels of frequency of the construction.

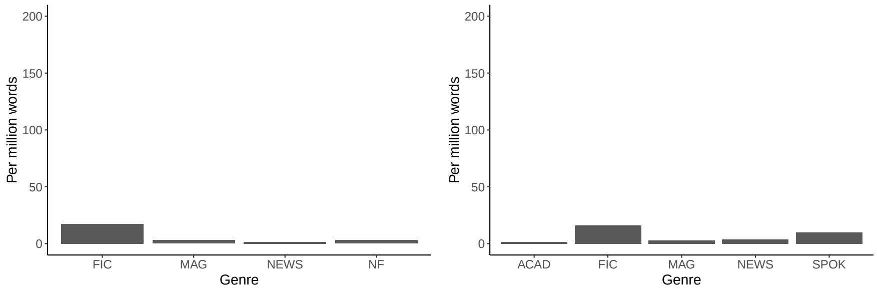

Figure 6.18: Text-linguistic frequencies of all occurrences of negative concord across corpus sections (COHA: n = 3,912; COCA: n = 2,917)

Negative concord is a very rare feature in edited standard American English. In this respect, it is fairly similar to *ain't*, with the difference that the frequency of *ain't* is higher in fiction than that of negative concord. The results are not surprising, given that the feature indeed disappeared from standard English during the seventeenth century (Nevalainen 2000; Tieken-Boon van Ostade 2008a; see also Nevalainen and Raumolin-Bunberg 2003), but remained a feature of the vernacular in both British and American English. What is interesting, however, is the significance of these results in

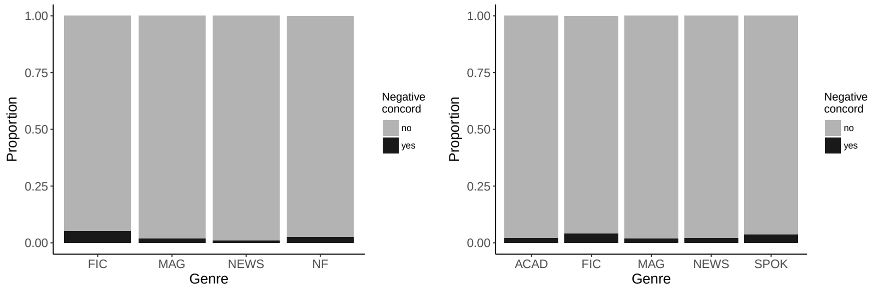

Figure 6.19: Proportion of negative concord (COHA: n = 3,912; COCA: n = 2,197) across corpus sections out of the total number of environments for negation with the indefinities *anybody*, *anyone*, *anything* (COHA: n = 91,165; COCA: n = 91,436)

the context of the treatment of this feature in usage guides. As I discussed in Section 5.2 above, negative concord is one of the features which is least frequently covered in the usage guides consulted. This may indicate that its frequency of occurrence is low in standard American English, and it is consequently not seen as a usage problem. Furthermore, the case of negative concord may provide evidence for the relationship between usage guides and frequency of use. As Ilson (1985) observed, a usage problem is usually a linguistic variant which has a high enough frequency of occurrence in order to be salient enough to be a usage problem. The reverse process might be taking place in the case of negative concord: the less the feature is used in standard American English, the less it will be treated in usage guides. On this basis, we could possibly even predict that negative concord is on its way out of the usage problem canon.

6.6 Pronouns in coordinated phrases

The proscribed forms of pronouns in coordinate phrases are also fairly low in frequency. On the basis of text-linguistic frequency, object *I* is somewhat less frequent than subject *me*, as shown in Figures 6.20 and 6.21, respectively; this difference, however, is not large, as the fluctuations in frequency for both features do not exceed 2.5 occurrences per million words.

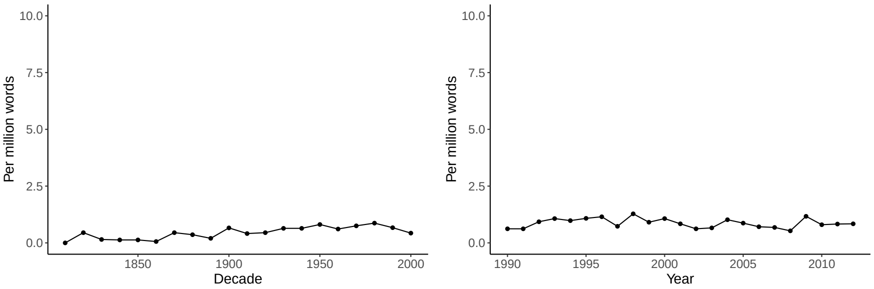

Figure 6.20: Text-linguistic frequency of object *I* across time (COHA: n = 194; COCA: n = 380)

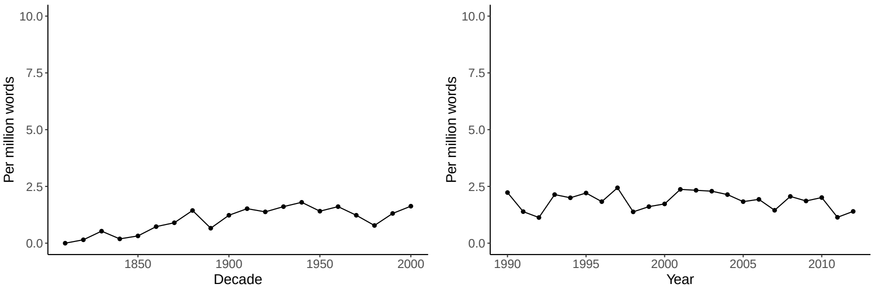

Figure 6.21: Text-linguistic frequency of subject *me* across time (COHA: n = 456; COCA: n = 819)

The results from the variationist analysis, given in Figures 6.22 and 6.23, partly support the text-linguistic frequency patterns. The proportions of both variants in relation to their standard counterparts are fairly low in the two corpora. There is one difference here with respect to the results from the text-linguistic analysis. While on the basis of the text-linguistic frequency distributions the occurrence of subject *me* is slightly higher than that of object *I*, especially in COCA, the variationist analysis shows that object *I* is more frequent than subject *me* when we take into account the total number of possible environments of each of the variants. This may indicate that while neither variant is very frequent in standard American English in terms of rate of occurrence, subject *me* is less often used in all possible environments compared to object *I* because it is seen as a more serious mistake. Object *I* is a well-known case of hypercorrection, and is considered a mark of formality. This difference between the two variants may account for the fact that the variationist analysis shows that object *I* is used more often than subject *me*.

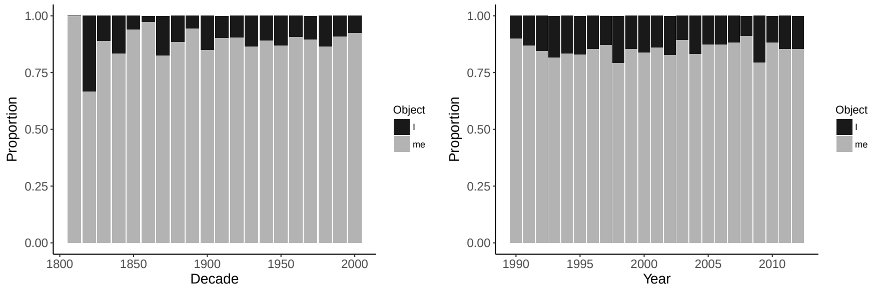

Figure 6.22: Proportion of object *I* (COHA: n = 194; COCA: n = 380) out of all possible environments across time (COHA: n = 1,808; COCA: n = 2,621)

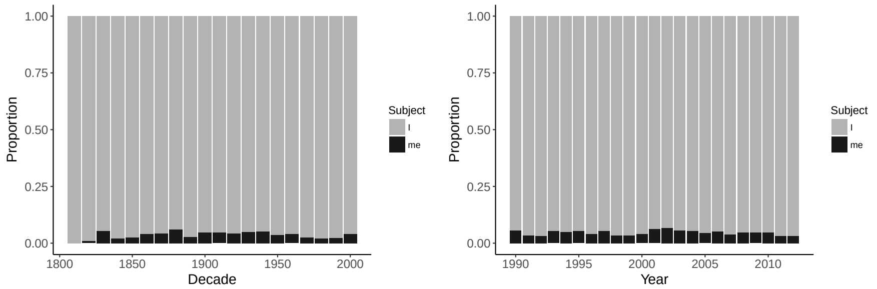

Figure 6.23: Proportion of subject *me* (COHA: n = 456; COCA: n = 819) out of all possible environments across time (COHA: n = 12,087; COCA: n = 17,546)

In addition to these analyses, I carried out further analysis on a portion of this dataset, focusing on cases of pronouns in coordinated phrases headed by *between*. The analysis consisted of manually disambiguating between cases with *between x and I* and cases with *between x and me*. It is important to note here that while the analysis based on all cases of object *I* and subject *me* were restricted to cases where the pronouns are used with a proper noun, the analysis of cases of *between x and I* and *between x and me* was carried out on the basis of all occurrences of the phrase, not only those with proper nouns. The phrase *between you and I* is the most commonly mentioned one in the entries on object *I*; consequently, the analysis considered this specific case in more detail, and explored the extent to which observations about this feature made in usage guides relate to patterns of actual use. In addition, this manual analysis was done in order to gain more reliable insights into the distribution of this feature, which is not possible to the same extent with automatic disambiguation. The dataset analysed is small enough for variants to be manually disambiguated, enabling us to gain an insight into the frequency distribution of one particular proscribed variant, which features strongly in discussions of object *I*.

The results from this analysis are presented in Figure 6.24. The frequency distribution shows that the variant *between x and I* is very infrequent. This in turn suggests that objections to the use of *between x and I* identified in the usage guides analysed do not relate to any evidence that the phrase is used frequently. The COHA data show that the proscribed variant is barely found during the nineteenth and twentieth centuries. The proportion of uses of *between x and I* out of the total number of possible environments is slightly higher in COCA, but this is a far from striking difference. On the whole, then, the variant is very infrequently used.

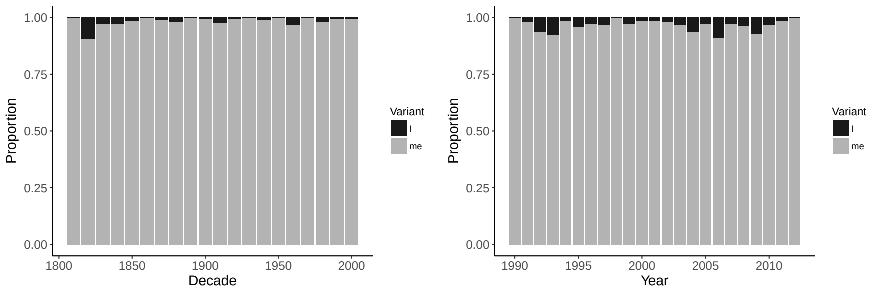

Figure 6.24: Proportion of object *I* and *me* in cases with *between* across time (COHA: n = 27; COCA: n = 44)

Turning to the distribution of object *I* and subject *me* across corpus sections, the pattern which emerges with respect to object *I* is expected. While the frequencies are overall very low, the COHA data have a slightly higher frequency of object *I* in fiction, while the spoken section in the COCA data contains most cases of object *I*, followed by fiction. This shows both that object *I* is infrequently found in general American English, and that when it is used, it is restricted to spoken registers. Of the written registers, fiction comes closer to colloquial text types, so these results are not surprising. As for subject *me*, the pattern of frequency distribution is similar to that of object *I* in the data from COHA, with fiction texts containing the highest rate of occurrence of subject *me*. In the COCA data, however, the situation is more striking. While object *I* seems to be most common in spoken texts, followed by fiction, subject *me* is most often used in fiction, while its use in the spoken sections is not higher than that in magazines or newspapers. This could perhaps be explained in part by the composition of the corpus sections in more detail. The spoken section of COCA contains spoken texts taken from television programmes, which means that

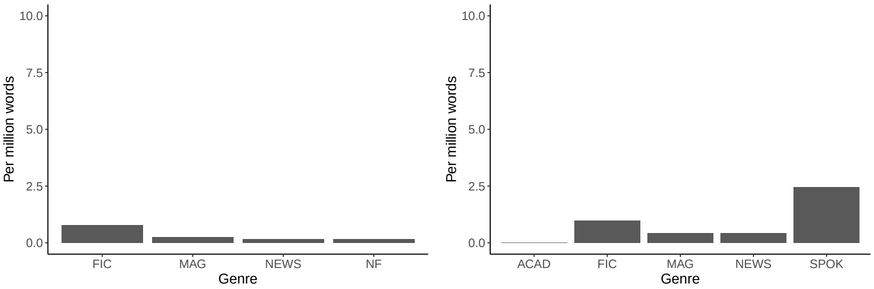

Figure 6.25: Text-linguistic frequency of object *I* across corpus section (COHA: n = 194; COCA: n = 380)

the language found in this section may not be as colloquial and informal as what one would expect to encounter in everyday colloquial settings. More specifically, when it comes to proscribed variants such as subject *me*, speakers in these contexts may have a tendency to avoid such uses altogether, which might explain the relatively low frequency of subject *me*. Fiction, on the other hand, contains a fair number of film scripts alongside novels and other fiction texts, such as short stories. The language in film scripts can be expected to be affected less by prescriptive norms than the language used by speakers in at least some television programmes. This might account for the higher frequency of use of subject *me*. While this may explain the distribution of subject *me* across corpus sections, it does not provide a satisfactory explanation for the difference in the patterns of occurrence of object *I* and subject *me*. This is, I believe, related to the difference in the features themselves. The fact that object *I* is more frequent in standard spoken data may suggest that it is indeed seen as a less serious error than subject *me*, whereas subject *me* is considered to be characteristic of very informal colloquial language use, and is consequently more frequent in fiction texts, including movie scripts.

The variationist analysis of object *I* and subject *me* reveals that while the text-frequencies of object *I* (Figure 6.25) are lower than those of subject *me* (Figure 6.26), the situation is reversed when we look at the proportion of uses of the two variants out of the total number of possible environments. While the prescriptively targeted variant object *I* appears to be most frequent in the spoken sections of COCA, it is also relatively frequent in the academic section, which might be seen as unexpected, given that academic texts are usually heavily edited, and proscribed variants would be expected to be rare (Figure 6.27). The same goes for its distribution across sections in COHA, where the magazine, newspaper, and non-fiction sections

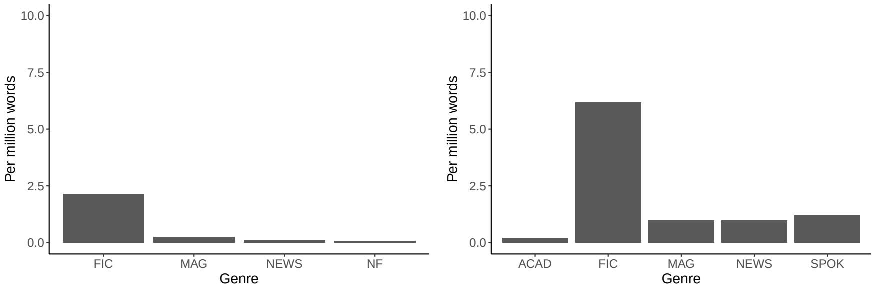

Figure 6.26: Text-linguistic frequency of subject *me* across corpus section (COHA: n = 456; COCA: n = 819)

contain the highest rates of object *I*.

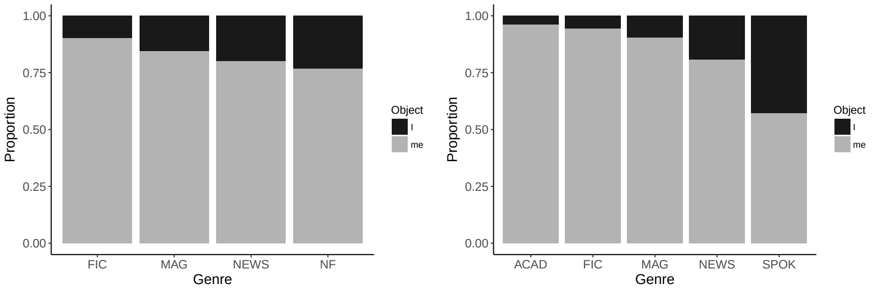

Figure 6.27: Proportion of object *I* (COHA: n = 194; COCA: n = 380) out of all possible environments across time (COHA: n = 1,808; COCA: n = 2,621)

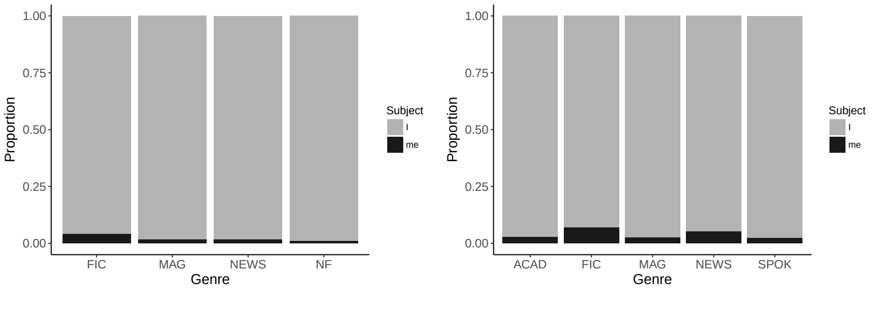

Figure 6.28: Proportion of subject *me* (COHA: n = 456; COCA: n = 819) out of all possible environments across corpus section (COHA: n = 12,087; COCA: n = 17,546)

Finally, the results from the variationist analysis of *between x and I* across sections of the corpora, shown in Figure 6.29, indicate that *between x and I* is most commonly found in the spoken and fiction sections. However, the proportion of the uses is still relatively low.

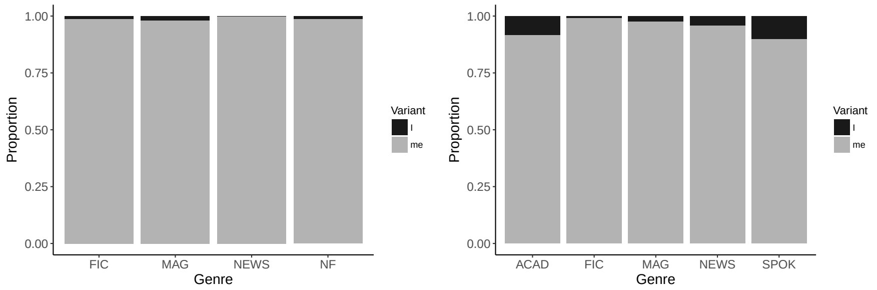

Figure 6.29: Proportion of *between x and I* (COHA: n = 27; COCA: n = 44) out of the total possible environments (COHA: n = 2,027; COCA: n = 1,362) across corpus section

In summary, both object *I* and subject *me* are very low in frequency in both COHA and COCA. In the data presented, we do not see any clear evidence of change in usage over time. This is also the case with the special case of this feature, *between x and I/me*. While both variants are very infrequent across time as well as across corpus sections, there is an important difference between the patterns of occurrence across time and across corpus sections. With respect to the former, the frequencies of both object *I* and subject *me* are very low, and there are no discernible patterns of change across time. It is important to note that the stability of the frequencies over time also indicates that the variants are not disappearing from the language. Furthermore, if the low frequencies of the features are in part a consequence of the fact that the corpora represent relatively standard language, it can be assumed that both variants are more frequent features of spoken language. With respect to the patterns of occurrence of the variants across sections of the corpora, the evidence suggests that object *I* is more often used in more standard or more formal colloquial registers, while subject *me* is more often used in more informal colloquial registers.

In the context of the coverage and treatment of these variants in usage guides, it seems that object *I* and subject *me* are rather straightforward cases of ‘old chestnuts’. The stability of their frequencies suggests that they are rare, but possible variants, and are mostly used in informal colloquial speech. The fact that they continue to be included in usage guides indicates that they are still considered problematic, which explains the high number of RESTRICTED entries for object *I* and UNACCEPTABLE entries for subject *me*. In other words, usage guides are not reacting to an increase in frequency of usage of these variants. Rather, they may be reacting to register variation in the use of these variants, or they may simply be perpetuating prescriptions on the basis of an established tradition, in a way which does not consider contemporary evidence from actual language use. In part, this could be considered evidence that usage guide writers indeed do not always distinguish between spoken and written levels of usage.

6.7 The split infinitive

In this section I present the analysis of the split infinitive, the final feature investigated on the basis of text-linguistic and variationist frequencies. There is a difference in the way these two frequencies were calculated. The text-linguistic frequencies were calculated on the basis of the identification of all infinitives split by one word, including *-ly* adverbs, other types of adverbs, and the negator *not*. The variationist frequencies were calculated on the basis of identifying the variable MODIFIED INFINITIVE, which is defined for the purposes of this analysis as any full infinitive modified by a single *-ly* adverb.

The text-linguistic frequency of split infinitives across time is given in Figure 6.30. The data show clearly that the rate of occurrence of the split infinitive has indeed been undergoing an increase; this is especially clear for the COCA data. There is a sharp drop in the trend for the last decade in the COHA data, which is surprising, and rather difficult to explain, because there is not a similar drop in the same decade in COCA. This might be in part a result of the fact that COCA and COHA have a different make-up (COHA is composed of about 50% fiction texts). What is important

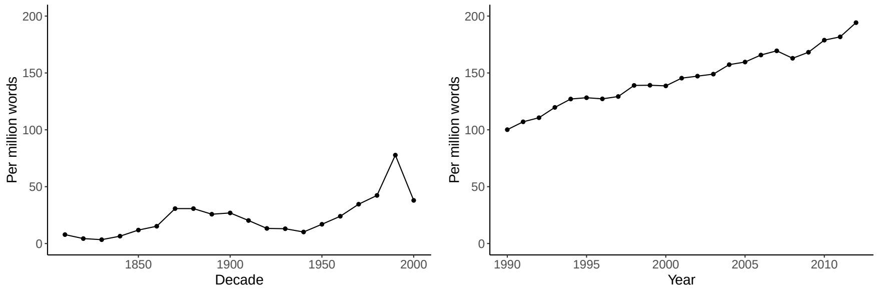

Figure 6.30: Text-linguistic frequency of split infinitives across time (COHA: n = 10,062; COCA: n = 63,079)

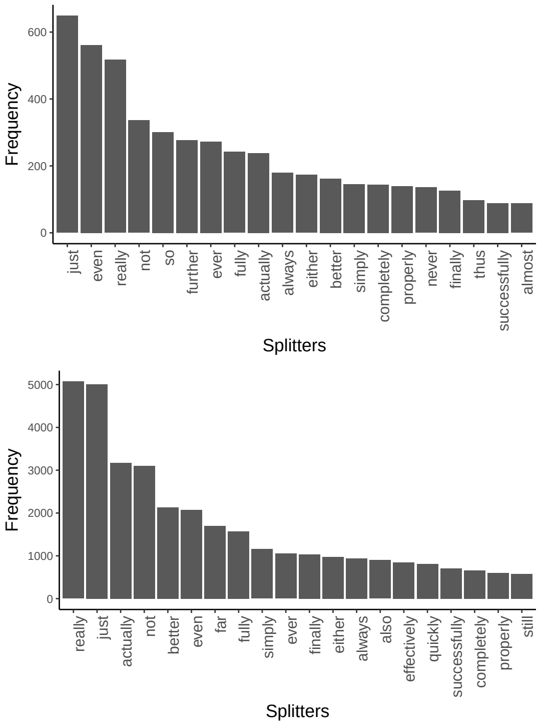

Figure 6.31: Most common splitters in COHA and COCA

though, as I will show below, is that this drop in the data disappears when we take a variationist approach to this feature (see Figure 6.32). This example nicely illustrates the point made by Biber et al. (2016) that normalised text-linguistic frequencies and variationist frequencies often produce differing accounts of the use of a particular variant. On the basis of this, I think it is not unreasonable to assume that the low rate of occurrence of the split infinitive in the last decade of the COHA data may be the result of the types of materials included in the corpus.

In addition to plotting the text-linguistic frequencies of occurrence of the split infinitive, we can perform an analysis on the items which most commonly split infinitives, i.e. the so-called 'splitters'. Figure 6.31 shows that while lexical *-ly* adverbs are the most common splitters, other types of adverbs and the negator *not* are also very common. While it would certainly be of interest to explore all the potential constraints on the occurrence of the split infinitive, including the variation in the use of all splitters, for the present study I limited myself to analysing the proportion of split infinitives (out of the total number of modified infinitives) only in contexts where the modifier is a lexical *-ly* adverb, as explained in more detail in Section 4.4.

The results from the analysis of the occurrence of split infinitives from the variationist analysis corroborate the increase observed in the text-linguistic frequency of the split infinitive. As Figure 6.32 shows, there is a definite increase in split infinitives over time, though the increase is only small during the second half of the nineteenth century, and is matched by a similar increase in the text-linguistic frequencies. After the middle of the nineteenth century, the trend decreases, and it picks up again after the 1940s. Since then, there has been a steady increase in the use of this feature.

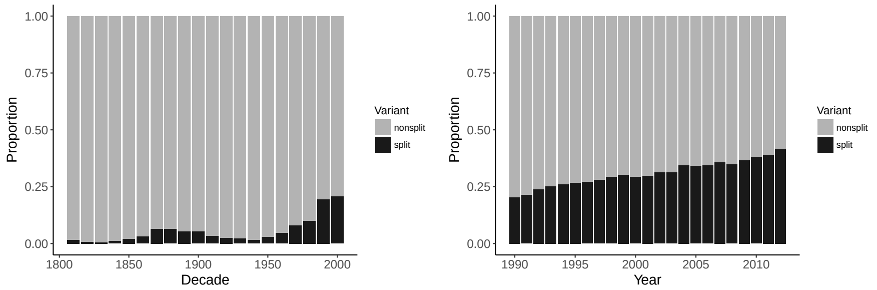

Figure 6.32: Proportion of split infinitives with lexical *-ly* adverbs across time (COHA: n = 6,037; COCA: n = 40,053) out of the total number of modified infinitives (COHA: n = 108,399; COCA: n = 130,855)

The use of the split infinitive across genre sections of the corpora (Figures 6.33 and 6.34) reveals that in the COHA data the rate is much lower; the newspapers section seems to have a slightly higher frequency of split infinitives, but on the whole the frequencies of occurrence of split infinitives in all sections in COHA are low compared to those in COCA. In the data from COCA, the rate of use of the split infinitive varies across sections, with spoken texts containing the highest rate of split infinitives, followed by academic. Magazines and newspapers have more or less equal number of occurrences of split infinitives per million words, while fiction has the lowest frequency of all sections.

The variationist analysis of the proportion of infinitives split by a single *-ly* adverb as opposed to non-split infinitives across corpus sections in COHA and COCA is plotted in Figure 6.34. The proportions in the figure exhibit similar patterns to those based on text-linguistic frequencies, which confirms the text-type distribution of split

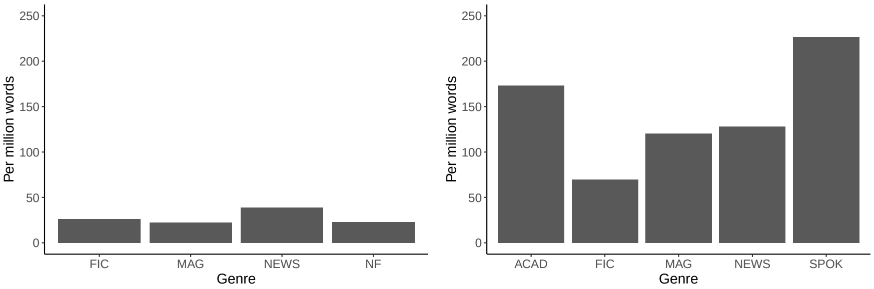

Figure 6.33: Text-linguistic frequency of split infinitives across corpus section (COHA: n = 10,062; COCA: n = 63,079)

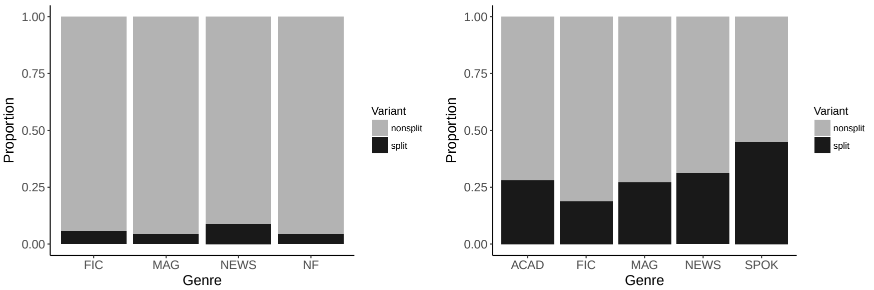

Figure 6.34: Proportion of split infinitives with lexical *-ly* adverbs across corpus sections (COHA: n = 6,037; COCA: n = 40,053) out of the total number of modified infinitives (COHA: n = 108,399; COCA: n = 130,855)

infinitives, i.e. the fact that they are most commonly found in speech. Academic texts show a different pattern, however. While the text-linguistic frequency of split infinitives is higher in academic texts than in magazines and newspapers, the proportional frequencies shown in Figure 6.34 are more or less the same for all three sections.

Turning to the importance of these results for the question of how these trends relate to the usage guide treatment of the split infinitive, as well as the changes observed in that treatment, this case presents us with two possible scenarios. First, on the basis of the increase in the use of the split infinitive after the 1950s, it might be argued that the split infinitive has increased despite prescriptive pressures against its use. This is an observation which has been made in previous studies (e.g. Calle-Martín and Miranda-García 2009; Leech et al. 2009). However, the analysis of treatment of the split infinitive discussed in Section 5.3 suggests that the treatment itself has started to change, and is becoming more accepting of the split infinitive. In fact, Albakry (2007) has also shown, on the basis of a smaller set of usage data and style guides, that, compared to the other usage features he looked at (sentence-initial coordinating conjunctions, stranded prepositions, functional shift, and modifying absolute adjectives), the split infinitive is not a strongly dispreferred feature. This brings me to the second scenario, in which we might consider the increase in the use of the split infinitive to be a consequence of the loosening of the stricture against its use.

The problem remains, however, of the impossibility of explaining this kind of increase in the use of split infinitives in terms of a weakening of prescriptive influence only. What we can observe here are two separate trends: one, in prescriptive literature, of loosening the prescription against the split infinitive, and the other, in the actual usage observed here, of increasing patterns of use of the feature. Again, as in other cases, this can be interpreted in three ways: first, prescriptivism influences usage; second, prescriptivism is influenced by usage; and third, there is no connection between these two whatsoever, and the observed change is coincidental. In addition to these three possibilities, it is important to consider a fourth one, which is a combination of the three possibilities, which are not mutually exclusive. However, in order to investigate this case further, and in the hope of gaining a better understanding of the level at which prescriptivism might affect language use, I conducted a multifactorial analysis to explore the extent to which the use of other proscribed features in a text may predict the use of one proscribed feature.

6.8 Identifying prescriptive influence at the textual level

Having explored the evidence for potential prescriptive influence, and having applied the traditional approaches in interpreting the trends observed, I now turn to a different approach to investigating prescriptive influence.[3] In the preceding sections, I explored the patterns of use of the six linguistic features investigated in the study, in order to gain insights into how they are used, with the ultimate goal of shedding light on the relationship between usage guides and actual language use. I applied both text-linguistic and variationist metrics in order to obtain more robust evidence for the patterns of use of the linguistic features investigated. While this approach revealed interesting and relevant aspects of the relationship between usage guide

[3] A version of this section also appears in Kostadinova (forthcoming).

treatment of usage and patterns of actual use, it still presents us with the challenge of ascertaining prescriptive influence in a manner that goes beyond the difficulty of equating correlation with causation. We also saw that prescriptive influence, even if it exists, is crucially conditioned by a number of other aspects of language use, both linguistic and stylistic, as the innovative study by Hinrichs et al. (2015) has shown. Inspired by their approach to the analysis of prescriptive influence in the use of restrictive relativiser *that*, I adopt and expand this approach by applying it to the analysis of the potential influence of prescriptivism on the use of the split infinitive.

The logic of this approach, as outlined in Section 4.4, is that many details of the variation patterns of a particular variant are lost when we look at corpus sections in terms of time periods or types of texts. Often, choices in usage which may be affected by prescriptivism are made by individual speakers or writers. So, while corpus-based frequency patterns might not on the whole contain any indication of potential influence of prescriptivism, this influence may be more readily identified at the level of individual texts. The level of specific texts thus provides a higher level of resolution at which prescriptive influence can be investigated.[4] The first assumption of this approach, then, is that prescriptive influence can be more meaningfully explored at the level of individual texts. The second assumption is that, if individual texts are influenced by prescriptive concerns for norms and correctness, this will be manifested in the use of many prescriptively targeted features simultaneously, not just one. Applied to the case of the split infinitive, I formulated the following hypothesis: if the split infinitive is influenced by prescriptivism in individual texts, the likelihood that a modified infinitive will be split will be higher in the presence of other prescriptively targeted variants.

There are a number of motivations for choosing to apply this approach to the split infinitive as a case study. The main motivation is the difficulty of arriving at a more decisive understanding of how prescriptivism has or has not affected the use of the split infinitive on the basis of the comparison between precept trends and actual use data discussed at the end of Section 6.7 above. The split infinitive is one of the features which seems to be losing its usage problem status, raising the question of the extent to which this has or has not influenced its use. Pragmatic motivations for focusing on the case of the split infinitive included the nature of the variable, as well as the size

[4] I use “texts” here to refer specifically to segments of language use included in the Corpus of Contemporary American English. In one sense, this is a specific use of the term, because it refers to corpus texts; in another sense, I use the term broadly, to refer both to more traditional types of texts, such as magazine articles, and to language segments which are not traditionally thought of as texts, such as television shows.

of the dataset. With respect to the former, the variants of the split infinitive are fairly straightforward to determine, and are both used widely in all kinds of texts. *Literally*, for instance, was not considered a good candidate for this kind of analysis, because of the difficulty of applying a variationist approach when analysing this feature. *Ain't*, on the other hand, was found to be restricted to specific text types. Similar issues were present for the other features included in this study.

The dataset for the analysis included all cases in which a *to* infinitive is modified by a single *-ly* adverb (see Section 4.4 for an explanation of how the dataset was produced). Each occurrence of a MODIFIED INFINITIVE was classified as either SPLIT, if the *-ly* adverb is placed between *to* and the verb, or NON-SPLIT, if the adverb is placed either before *to* or after the verb. Thus, the realisation of the variant SPLIT as opposed to NON-SPLIT modified infinitives was modelled as a binary choice in a binomial logistic regression model, the selection of which is explained in the next paragraph. Each case of a modified infinitive was classified as either SPLIT or NON-SPLIT; this was the dependent variable. A number of predictors were used in the model, as explained in Section 4.4, including internal predictors, ADVERB TYPE and ADVERB LENGTH; external predictors, YEAR and GENRE, and a number of prescriptivism-related predictors (see Section 4.4 for a more detailed explanation and examples). These predictors are other prescriptively targeted features, whose frequencies of occurrence in each individual text in the corpus were included as predictors in the model. The following language features were used in the model as prescriptivism related predictors: *ain't*, sentence-initial *and/but*, singular *data*, *hopefully*, *these kind/sort of*, plural *less*, the discourse particle *like*, *literally*, negative concord, plural *none*, passives, *shall*, *try and*, and *whom*. For each text in the corpus, I calculated the normalised frequency of occurrence per 1000 words for each of these features (see Appendix C on the extraction of these features from COCA).

The statistical model used to explore the relationship between the occurrence of split infinitives in a text in relation to other prescriptively targeted features was a binomial logistic regression model. The analysis was conducted on the basis of a procedure outlined in Levshina (2015). The best model was selected using backward stepwise selection on the basis of the lowest AIC (Aikake Information Criterion) value. In addition, the function `drop1` was used to check which of the predictors contribute significantly to explaining the variance in the dependent variable. On the basis of both the backward stepwise selection process, and the results on the predictors which significantly contributed to explaining the variance in the data, the model given in Table 6.1 was selected. As the final model shows, a number of the prescriptivism-related predictors did not survive the model-fitting stage: *these kind/sort of*, plural *less*, *literally*, negative concord, plural *none*, and *try and*. Only the predictors which are significant in explaining the variance in the dependent variable were thus included in the final model. Even though this model did not show significant improvement in the concordance index C compared to a model with all predictors included, the simpler model was selected, and the value for C was considered acceptable (see Levshina 2015: 259). Following the procedure outlined in Levshina (2015: 274), bootstrap validation was applied to the model to check for over-fitting, using the function `validate()` in the package `rms` (Harrell 2018). The model was refitted 200 times, and the optimism scores were low for all the goodness-of-fit statistics, indicating that the model is satisfactory in accounting for the relationship between the variables.

I now turn to an examination of the results for each predictor in the model. Starting from the internal predictors, the model shows that both ADVERB TYPE and ADVERB LENGTH are significant predictors, indicated in Table 6.1 by asterisks. For ADVERB TYPE, the reference level is ADDITIVE-RESTRICTIVE adverbs. This means that the results displayed in Table 6.1 show how the likelihood that a modified infinitive is split differs in cases in which an adverb is, for instance, a DEGREE adverb, as opposed to cases in which it is ADDITIVE-RESTRICTIVE. The results thus show that the likelihood that a modified infinitive is split is higher if the adverb belongs to one of the following four levels: DEGREE, MANNER, STANCE, or TIME, compared to cases in which the adverb is ADDITIVE-RESTRICTIVE. Cases with LINKING adverbs do not significantly predict the likelihood of a modified infinitive being SPLIT.

The second linguistic predictor, ADVERB LENGTH (measured in syllables), is also significant. As already explained in Section 4.4, ADVERB LENGTH is operationalised as the difference in number of syllables between the adverb and the verb in each case of a modified infinitive in the dataset; the variable has three levels: LONGER, if the adverb is longer than the verb; SHORTER, if the adverb is shorter than the verb; and EQUAL, if the adverb has the same number of syllables as the verb. The reference level here is EQUAL. Compared to cases in which the length of the adverb is the same as that of the verb, the odds of an infinitive being split decrease by 0.70 when the adverb is longer than the verb ($p < 0.01$). In other words, if an adverb is longer than the verb, it tends to come after the verb, rather than before. There was no such significant difference for shorter adverbs.

From the external predictors, I analysed YEAR and GENRE. The external predictor YEAR, which is operationalised as a continuous variable, and is associated with the year of publication of the corpus text in which each case of a modified infinitive was identified, is also significant, and shows that the odds of an infinitive being SPLIT

predictor:level	*b*	OR	*p*	
(Intercept)	−122.00	0.00	< 0.01	***
INTERNAL PREDICTORS				
adverb class:degree	2.07	7.88	< 0.01	***
adverb class:linking	1.00	2.72	0.05	
adverb class:manner	1.21	3.34	< 0.01	***
adverb class:stance	2.05	7.74	< 0.01	***
adverb class:time	1.57	4.82	< 0.01	***
adverb length:longer	−0.36	0.70	< 0.01	***
adverb length:shorter	0.14	1.15	0.33	
EXTERNAL PREDICTORS				
year	0.06	1.06	< 0.01	***
genre:fiction	−0.11	0.89	0.34	
genre:magazine	0.28	1.32	0.03	*
genre:newspaper	0.49	1.64	0.12	
genre:spoken	1.10	2.99	< 0.01	***
PRESCRIPTIVISM PREDICTORS				
ain't	0.46	1.58	0.08	
And/But	−0.03	0.97	0.02	*
data sg.	1.95	7.04	< 0.01	***
hopefully	0.89	2.43	0.09	
like	0.66	1.93	0.01	***
passives	0.04	1.04	< 0.01	***
shall	−0.80	0.45	< 0.01	***
whom	−0.80	0.45	< 0.01	***
SUMMARY STATISTICS		n = 4,925		
		LR χ^2	812.03	
		Pr(>χ^2)	<0.0001	
		df	20	
		R	0.205	
		C	0.729	
		Somer's Dxy	0.45	
observations	4925			
non-split	2873			
split	2053			

Table 6.1: Binomial logistic regression model for the alternation between SPLIT and NON-SPLIT infinitives modified by one *-ly* lexical adverb. Reference level is NON-SPLIT infinitive

increase by 0.06 for each one-unit increase in YEAR. The predictor GENRE is a categorical variable with five levels: ACADEMIC, FICTION, MAGAZINE, NEWSPAPER, and SPOKEN. The level ACADEMIC was used as the reference level in the model. The model shows that the likelihood of an infinitive being SPLIT is significantly different for the magazine section and the spoken section, compared to academic. Compared to academic, the likelihood that an infinitive is split increases by 1.32 ($p = 0.03$) in magazine texts. The significance is stronger for spoken texts: compared to academic texts, the likelihood that an infinitive will be split in spoken texts increases by 2.99 ($p < 0.01$).

Finally, the prescriptivism-related predictors show that the significant predictors here are: sentence-initial *and/but*, singular *data*, the discourse particle *like*, passives, *shall*, and *whom*. Of these, sentence-initial *and/but*, singular *data*, the discourse particle *like*, and passives significantly increase the likelihood of an infinitive being SPLIT. In other words, in texts in which these features occur, for every one-unit increase in the frequency of these features, measured as normalised frequency of occurrence of the relevant feature per 1,000 words, the likelihood of an infinitive being SPLIT increases. The statistical significance is the weakest for sentence-initial *and/but* ($p = 0.02$), while all other features are statistically significant predictors ($p < 0.01$). The highest increase in the odds that an infinitive is split is predicted by the occurrence of singular *data*; for each one-unit increase in the normalised frequency of singular *data*, the odds of an infinitive being split increase by 7.04 ($p < 0.01$). The other two significant prescriptivism-related predictors affect the likelihood of an infinitive being split in the opposite direction. For every one-unit increase in the normalised frequency of *shall*, the odds of an infinitive being split decrease by 0.45 ($p < 0.01$); the same result was obtained for *whom*. *Hopefully* and *ain't* are not significant predictors for the use of split infinitives.

Checking the model for interactions showed that the most interesting significant interaction is between YEAR and GENRE. Figure 6.35[5] shows the change in the odds of a modified infinitive being realised as split (as opposed to non-split) per one year for each GENRE level in the corpus separately. As evident from the plots, the change in odds across YEAR is different for the different GENRE levels. The figure shows that the odds that an infinitive is split decrease over the course of the period between 1990 and 2012 in the newspaper section of the corpus, while they increase in all other sections. Most interesting here is that the increase in the odds of a modified infinitive being realised as a split infinitive seems to be greatest in academic texts. There were a number of other interactions, but they did not produce any differences in direction, just in the size of the effect. Consequently, I will not discuss them here in further detail.

[5] This plot was produced using the `visreg` package in R (Breheny and Burchett 2017).

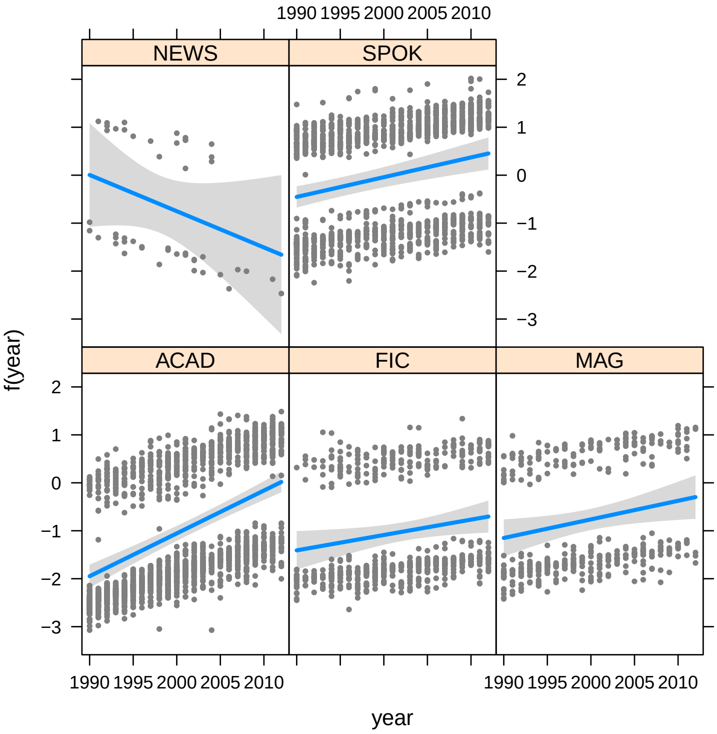

Figure 6.35: Interaction between genre and year

What do these results reveal about what constrains the use of split infinitives as opposed to non-split modified infinitives, and the role of prescriptive ideology in that context? As the binomial regression model shows, texts in which sentence-initial *and/but*, singular *data*, the discourse particle *like*, and passives are used would be less influenced by prescriptive strictures, and would consequently be more likely to contain split infinitives. In other words, writers or speakers who are not concerned about using, for instance, singular *data*, would also be unconcerned about using split infinitives. On the other hand, texts in which authors (or editors) use *shall* and *whom* would be texts in which split infinitives are less likely to be used. *Ain't* is not a significant predictor, because unlike all the other predictors, which belong to stylistic prescriptivism, *ain't* belongs to standardising prescriptivism (cf. Curzan 2014: Chapter 1). Thus, the choice of *ain't* over *be not* or *have not* forms is affected by a different set of considerations, which have to do with following a standard grammatical norm, rather than stylistic preferences. This of course does not apply to the case of *hopefully*, which Curzan (2014: 33–34) argues is an example of a feature related to stylistic prescriptivism, so in this case the explanation used for *ain't* does not hold for *hopefully*. It can be hypothesised that one possible reason that *hopefully* does not significantly predict the likelihood of an infinitive being split could be that *hopefully* itself is not affected by prescriptive ideology. In any event, this is something which remains to be further investigated in the future.

The interaction identified in the model provides further evidence of how likely the split infinitive is to be used in different types of texts. The most striking finding here is that the increase in the likelihood of an infinitive being split is greatest in academic texts; this suggests that the change towards more split infinitives is led by its use in academic language. Since the increase of the likelihood of an infinitive being split can also be identified for fiction, magazine, and spoken texts, it is reasonable to expect that split infinitives will increasingly be used in those types of texts as well. On the other hand, in newspaper texts, the odds of an infinitive being split decrease across the time period studied, as shown in Figure 6.35. An issue with relying too much on this finding is that the confidence intervals are fairly large, and the level NEWSPAPER was not significant in the model discussed above. Any interpretation would thus have to be made tentatively. This is an indication, albeit weak, that the newspaper genre might still be influenced by stylistic prescriptivism.

6.9 Conclusion

A number of observations can be made based on the results of my analysis of the six features separately, as well as on the results taken together. First, with respect to the six features separately, perhaps the most surprising result is the decrease in the frequency of use of *ain't*. The results do not bear out our original assumption that the increased acceptability of this feature will result in an increase in use. What seems to have happened is that the public discourse on *ain't* may have affected the frequency of use much more than the discussion of this form in usage guides. This case shows that the ways in which usage guides respond, if they do so at all, to ongoing changes in language use are different for different features. The other interesting case is the use of the discourse particle *like*, which is a clear-cut case of prescriptivism responding to a highly salient language change. In this case, it is highly unlikely that we will see a strong influence of prescriptivism on the use of this feature. *Literally* does not seem to be affected by prescriptivism either. The proportion of the non-primary uses of *literally* has increased over time. The case of *literally* also shows that the reactions to its non-literal use, which tend to exaggerate the frequency with which the feature is actually found, do not seem to be based on empirical evidence. The use of negative concord is a stable non-standard feature, and here there is little change both in how it is used and how it is treated in usage guides. My analysis of the use of pronouns in coordinated phrases shows that object *I* and subject *me* are predominantly restricted to informal contexts, as the proportion of these variants in the corpus data was not very high. The corpus data provided some interesting evidence that the notion that object *I* and subject *me* are problematic in a different way, as shown by the difference in their treatment in usage guides (see Section 5.3), may be borne out by corpus evidence. Finally, the split infinitive is a complicated case, which presents us with the difficulty of ascertaining prescriptive influence by relying solely on a comparison between precept and practice. This kind of comparison for the split infinitive confronted us with more than one possible interpretation of what may be the case in reality. The novel approach applied to the analysis of multiple possible factors constraining the use of the split infinitive showed that the split infinitive is a stylistically prescriptive feature which seems to be favoured in some cases and disfavoured in others. Academic texts, which perhaps tend to be less stringent when it comes to stylistic prescriptivism, seem to be promoting the change towards split infinitives. Other text types, however, may not follow the same trend. While more research certainly needs to be done for this finding to be corroborated for other text types, I believe the results show the complex and dialectic nature of the interplay between prescriptivism and actual use. In other words, in the long run split infinitives may certainly be expected to continue to be used (and critics of prescriptivist efforts may use this case as yet another example of the failure of prescriptivism). However, at present the use of the split infinitive may still be constrained by prescriptivist concerns, and this may be especially true in the context of specific text types.

All in all, the results show the complicated nature of the relationship between prescriptivism and actual usage, which prevents us from making generalisations based on individual features alone. It appears that for some features, such as *ain't* and negative concord, prescriptivism may have an influence over a longer period of time, but these features are non-standard and highly stigmatised. Even in these cases, the usage guide tradition alone may not have a strong influence if it is not backed up by a public discourse denouncing these features, as well as the educational system, through which non-standard features are regulated. In these cases, we see an example of what

Curzan (2014) calls standardising prescriptivism. In the case of the split infinitive, which is a typical stylistic feature, the influence of prescriptivism is of a different nature, and may be restricted to individual cases – speakers or texts – but not at the level of the language system. Finally, the case of the discourse particle *like* is perhaps the most recent example of how prescriptivism can respond negatively to changes in the language, which is one of its most distinguishing characteristics. Even though it is questionable whether over time prescriptivism will have an effect on the use of the discourse particle *like*, this remains to be seen.