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

Kostadinova, V.

CHAPTER 4

Methodology

4.1 Introduction

As outlined at the end of Chapter 1, the main purpose of this study is to empirically investigate present-day prescriptivism in American English by examining a small set of language features which are generally known to be usage problems for ordinary speakers. This investigation will be carried out by approaching these features from three perspectives: (a) attitudes to these features found in American usage guides, (b) the actual patterns of variation and change in the use of these features in corpus data, and (c) ordinary speakers' attitudes towards the use of these features. In this chapter I outline the approach taken in this study, first by discussing my general approach to the study of prescriptivism in Section 4.2, and then by presenting the types of data and analysis used for the present study in Sections 4.3, 4.4, and 4.5. The study is not limited to one particular language feature; rather, it focuses on a small set of features, in order to explore in detail the extent to which different features are affected differently by prescriptivism. The selection of the features was discussed in Section 3.2. The different perspectives I will adopt in analysing the features are explored by analysing three types of data. The metalinguistic treatment of the features in usage guides will be studied on the basis of an analysis of entries on the selected usage

features found in American usage guides. The patterns of variation and change of the selected language features will be analysed using American English corpus data, and a range of methods, including a quantitative multivariate analysis of frequency patterns and constraints on the distribution of different variants in the context of one of the features described in Chapter 3, the split infinitive. Finally, speakers' attitudes will be explored with data from a survey and post-survey interviews conducted with native speakers of American English.

4.2 General approach

The overall approach to investigating present-day language prescriptivism in American English taken in this study, as mentioned in the introductory chapter above, is three-pronged; it aims to explore the three main perspectives and to account for the potential influence of prescriptivism on language use. This tripartite division is reflected in the methodology and the data used. Generally, the study applies both qualitative and quantitative methods to the analysis of the three types of data mentioned above. There are a number of ways in which the present methodological approach alleviates some of the problems and difficulties in studying prescriptivism established in previous research.

The first aspect of the approach is that the study is significantly informed by the comparison of precept and practice (Konopka 1996, cited in Auer 2009; Gustafsson 2002; Auer and González-Díaz 2005), a methodological approach often taken in studies of the influence of prescriptivism and normative linguistics on language change. As discussed in the theoretical background presented in Chapter 2 above, prescriptivism is sometimes assumed to have an influence on language, but in many of the cases where such an influence is assumed, prescriptivism is taken to be a static normative phenomenon, which is often approached as a phenomenon which does not change over time. However, numerous discussions of prescriptive ideology have shown that prescriptivism does in fact change over time, so in order for prescriptive influence to be ascertained reliably, an analysis of prescriptive ideology and institutionalised prescriptive attitudes needs to be carried out (cf. Curzan 2014).

The second aspect of the approach taken here is that I focus on multiple language features, rather than on a single feature (e.g. Auer 2006; Wild 2010; Yáñez-Bouza 2015) or a set of related features (e.g. Anderwald 2012, 2016). In terms of language features, I focus on a small set of linguistically unrelated features, described in

Chapter 3 (for the approach to analysing their variation in the context of their respective variables, see Section 4.4), but I also account for their relationship to other prescriptively targeted features (see further Section 4.4). The aim here is to find a middle ground between the detailed, exhaustive approach taken in previous studies focusing on individual language features, and a more overarching approach that attempts to account for the way in which different language features are affected by prescriptivism.

Thirdly, the study covers the period from the middle of the nineteenth century to the present day, with special emphasis on present-day American English. In other words, I look at prescriptivism during the prescription stage of the process of standardisation, a little-studied stage in the context of studies of prescriptivism and its effects. Previous research on the topic is biased towards the stage of codification and has tended to focus on the eighteenth (e.g. Auer and González-Díaz 2005; Yáñez-Bouza 2015) and nineteenth (e.g. Dekeyser 1975; Anderwald 2014) centuries. The data on prescriptive attitudes to usage, as well as on language variation and change, go further back in time, to the middle of the nineteenth century, in order to track potential changes in prescriptive attitudes. Furthermore, Auer (2009: 9) has pointed out the need for language corpora to cover longer periods of time than the precept corpora (i.e. the collection of metalinguistic, or precept, data). In this way, any identified changes in precept may be tested for their influence by looking at the language use before and after the period in which changes in precept might have taken place, allowing for a time gap for the potential influence to be reflected in language use data.

Lastly, I include data on speakers' attitudes; in the discussion of the relationship between speakers' attitudes and prescriptivism the focus will be on present-day American English. As already mentioned in Chapter 1, speakers' attitudes are crucial to a better understanding of the social influence of prescriptive ideology. In historical sociolinguistic studies, such data are fairly difficult, if not impossible, to obtain. On the other hand, in surveys of attitudes to usage conducted in the course of the twentieth century (such as Leonard 1932; Marckwardt and Walcott 1938; Crisp 1971; Mittins et al. 1970), the attitudes to usage investigated are those of language professionals, not ordinary language speakers. In dealing with the attitudes of ordinary language speakers, this study aims to contribute to our understanding of prescriptive influence and attitudes to usage. The inclusion of an analysis of speakers' attitudes has been one of the critical aspects of the broader research project which this study is a part of, as mentioned in Chapter 1.

The precept *vs.* practice approach is, of course, not without problems. While the analysis of precept has been shown to be indispensable to the study of prescriptive influence, the relationship between changes in precept and changes in practice is problematic. As discussed in Section 2.3, a few studies based on the precept *vs.* practice approach have found limited influence of prescriptivism on language variation. Such studies, however, come with an important caveat, namely that correlation is not necessarily causation (Hinrichs et al. 2015), and, because of this, Curzan (2014: 84) observes that conclusions based on relationships between prescriptivist judgements and corpus evidence should be drawn very carefully. In order to deal with this limitation, the methods used in this study will go beyond the comparison of precept and practice patterns, and will operationalise the analysis of prescriptive influence by looking at the co-occurrence patterns of a number of prescriptively targeted features, alongside the six features which are at the centre of this study. This approach is based on Hinrichs et al. (2015), who investigated the extent to which the use of *that* as opposed to *which* correlates with the use of other prescriptively targeted features such as split infinitives, passives, sentence-final prepositions, and future reference *shall* with first person subjects, on the basis of data from the BROWN family of corpora. This kind of approach allows for a more thorough investigation of the possible prescriptive influence on the use of specific variants. These additional features are discussed in detail in Section 4.4.

4.3 Usage guides: data and analysis

The analysis of attitudes to usage in American usage guides is based on an analysis of 70 guides published in America between 1847 and 2014 (see Primary Sources). Entries on the language features investigated here were collected and analysed across a number of dimensions, discussed below. The selection of the usage guides used in the analysis was carried out in large part on the basis of the HUGE database, described in Section 3.2. A search for American usage guides yielded 44 guides, of which eight are classified as both British and American in HUGE. The reason for the double classification of these usage guides is that they cover linguistic features used in both British and American English. In the HUGE database user manual, Straaijer (2015: 5) notes that "[i]n cases in which the language variety was not explicitly mentioned, the variety was usually assigned based on the country of publication and the nationality of the author". However, a closer look at the doubly classified usage guides revealed

that while some of them do address both British and American standards of usage (a good example of this is Peters 2004), others point out in their prefaces that they are mainly concerned with British usage. Such is the case for Greenbaum and Whitcut (1988), for instance, where the authors explicitly point out in their preface that they "address [themselves] primarily to those wanting advice on standard British English" (1988: xiii). Four of these guides were thus excluded from the present analysis, resulting in the use of a subset of 40 American usage guides from the HUGE database. Given that the database contains both British and American usage guides, and is less comprehensive for the twentieth century than for the nineteenth (cf. Straaijer 2015), additional research was carried out to take into account more usage guides written for American English and published in the twentieth century. This additional selection was based on the criteria used in the selection of material for the HUGE database, in order to ensure consistency in the collection of guides, *viz.* selecting those guides that treat predominantly grammatical usage problems. A further criterion was the selection of only American usage guides. This additional search process consisted of searching the digital libraries HathiTrust and Internet Archive, Google Books, and the Leiden University Library Catalogue for additional titles. The additional search produced 30 usage guides, which were added to the 40 available in HUGE, thus bringing the total number of American usage guides consulted for the purposes of the present analysis to 70.[1]

The definition of what a usage guide is, as well as the delimitation of the genre of usage guides, is not a straightforward task, because of the variation in the form and content of these works (Straaijer 2018: 12–13), an issue I raised in Section 2.2. As I explained there, the question of the genre differences in these kinds of metalinguistic works has been raised elsewhere (e.g. Connors 1983; Weiner 1988; Straaijer 2018), but it has not always been consistently applied to analyses of usage guides. It is important to point out that very few of the studies of usage guides go into much detail on the selection of materials and the definition of usage guides, or other metalinguistic reference works, such as language manuals, handbooks, and textbooks. Meyers (1995) for instance focuses on handbooks of composition, but uses the terms textbooks and handbooks interchangeably to refer to the 60 books he analyses. No further details are included as to the selection of materials. A similar gap relating to

[1] In some cases, usage guides are published in both the United Kingdom and the United States. Such is the case, for instance, for Brians (2003). In the case of Partridge (1947), the first edition of his usage guide was annotated for American English, and these notes were given in square brackets in the first British edition.

the data selection process is present in Albakry (2007), who used 18 usage books to analyse the treatment of a number of features; the criteria for selecting these 18 books, however, are not explicitly addressed. Peters and Young (1997) similarly provide no explicit criteria for determining what a usage guide is. The compilation of the HUGE database dealt specifically with the question of delimiting the genre of usage guides, and establishing the criteria for what counts as a usage guide or not (cf. Straaijer 2018); these criteria were also applied in the present study.

It is also important to discuss the extent to which these 70 usage guides are representative of general usage guide publication trends in American English. While it is impossible to come up with exact figures for the total number of guides published in the United States, it is possible to arrive at an approximate picture of the publication trends. The reasons behind the difficulty in ascertaining exact figures are that (a) it is impossible to know for certain how many usage guide titles were published in total, and (b) an attempt to ascertain this total number of usage guides would also depend largely on one's definition of usage guide, which, as I have argued above, is difficult to establish (cf. Straaijer 2018). Thus, Straaijer (2018) notes that the estimated number of usage guides ever published may be between 250 and 300 titles, depending on one's definition (see also Tieken-Boon van Ostade forthcoming). Based on the definition and genre characteristics discussed in Straaijer (2018), and a number of additional sources consulted (e.g. the bibliographies of Gilman 1989 and Garner 1998), a list of the total number of usage guides published in America was produced. The publication trends are given in Figure 4.1. An important point to make is that the considerable variation in the genre means that these figures should not be taken to represent absolute numbers of published guides per decade, but should rather be interpreted as an approximation to the real situation. There may be guides that were published, but not identified in the process. In addition, if different selection criteria were applied, the results might be slightly different.

Figure 4.1 shows the general increase in usage guide publications during the twentieth century. While I noted that it is difficult to provide exact figures, similar trends have been observed elsewhere (Straaijer 2018; Tieken-Boon van Ostade forthcoming). In addition, earlier work on handbooks of composition, a genre which can be considered related to usage guides, has shown that these types of books were also on the increase. Meyers (1995: 30), for instance, analysed "60 handbooks of composition published between 1980 and 1993 (33 between 1990 and 1993 alone)". What is interesting here is that the number of handbooks is strikingly higher in the three-year period between 1990 and 1993. Even though Meyers does not provide

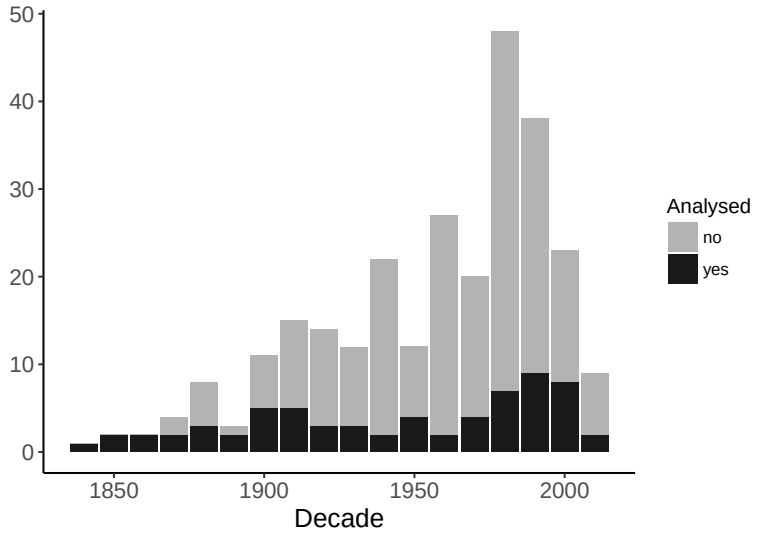

Figure 4.1: Total number of guides published in America per decade and the number of guides included in the analysis

details on the process of sampling of the analysed handbooks, or the extent to which his findings are representative of general publication trends, given the relatively high number of handbooks analysed, it might be safely concluded that the number of handbooks published grew significantly during the last decade of the twentieth century.

Figure 4.1 also shows that the number of usage guides analysed is not the same across all decades, which may be considered a limitation of the dataset. However, given the difficulty in obtaining many of these works, the inclusion of at least two usage guides per decade was considered to be adequate. When it comes to selecting a representative group of usage guides, the decision would not only depend on obtaining a representative number of guides, but also on selecting influential and popular guides. This aspect presents a different set of challenges, because determining the influence or popularity of a guide is not always straightforward. Despite these potential limitations, which will be kept in mind in the interpretation of the findings, the present collection of usage guides is significantly larger than those used in previous studies. The three most comprehensive studies of usage guides to date, for instance, analysed fewer usage guides: Creswell (1975) analysed ten usage guides, Peters and Young (1997) fourteen American usage guides, and Albakry (2007: 33) “a sample of the eighteen most popular usage books published in the United States since 1950”. In drawing on

a larger collection of usage guides, it is hoped that a clearer picture will emerge of the attitudes to usage and changes in those attitudes over time.

In most instances, the first edition of the usage guide in question was consulted, and the year of the first edition was used in the analysis of the usage guide data, as will be explained in more detail below. This also applies to subsequent impressions of the first edition. For fifteen out of the seventy selected usage guides, I was not able to obtain the first edition, nor was I able to ascertain the extent to which the editions that I consulted were changed compared to the first edition. This is a problem that has cropped up before in studies of normative grammars in historical sociolinguistics (see Wild 2010: 31, footnote 11). Yáñez-Bouza (2015: 29), for instance, argues for considering multiple editions of the texts in questions "on the assumption that different printings are likely to show modifications in the discussion of the same topic". On the other hand, with reference to eighteenth-century grammars, and specifically Lowth's grammar, Tieken-Boon van Ostade (2008b: 122) notes that new editions were sometimes advertised as corrected as a publisher's ploy to increase sales.

Although this discussion concerns eighteenth-century grammars, a similar case can be made for usage guides, which may be published in a 'revised' edition, but for which it may be rather difficult to ascertain to what extent the edition has changed, given that in most cases multiple editions of the same work were not available. For these usage guides, I decided to take the year of the edition consulted, as looking at the actual years of the revised editions showed that this may not be a critical issue in the analysis of the data. Since the data were analysed across decades, usage guides for which the edition consulted here was published in the same decade as the first edition were associated with the same time period in the analyses. This is the case with Witherspoon's *Common Errors in English and How to Avoid Them*, which was first published in 1943, while the edition I consulted was published in 1948; there were three such usage guides in total. Of the remaining twelve, most of the revised editions consulted were published within ten years of the first edition, so the difference in dating them was only one decade. In that case, the year of the revised edition was used, on the assumption that the time difference would not critically affect the analysis of the results; this was the case with eight of the usage guides. Finally, of the remaining four, the difference between the first and the consulted revised editions was two decades for three usage guides, and four decades for only one usage guide, Ebbitt and Ebbitt's *Writer's Guide and Index to English* (1978), first published in 1939.

As mentioned above, entries on the six selected usage features were identified and used in the analysis of attitudes to usage (which, as will be discussed below, are not to

be confused with speakers' attitudes). The HUGE database allows users to download entries on specific features in various formats, including xml. I used this file format to create a corpus of entries in which the text for each entry is a separate text file. This file is linked to another file which contains the metadata for each entry, such as author, title, year, and page numbers. Finally, for each entry, there was a third type of file where annotations were stored. For the usage guides which were selected additionally, the text, metadata, and annotation files for each entry were created manually. This resulted in a corpus of 281 entries in total for the six language features, given in Table 4.1.

Feature	**No. of entries**	**Average no. words per entry**
ain't	46	243.69
like	10	180.82
literally	32	173.09
negative concord	42	294.71
pronouns: object *I*	73	210.90
pronouns: subject *me*	19	347.42
split infinitive	59	381.58
total	281	268.57

Table 4.1: Number of entries across linguistic features in the corpus

On the basis of previous studies of normative grammars and usage guides outlined in Sections 2.2 and 2.3 above, a framework for analysis was created in an attempt to provide a comprehensive account of the treatment of the linguistic features under investigation and the attitudes expressed towards them. Previous studies have shown that a prescriptive approach to usage can be manifested in different ways in usage guides, including the manner in which opinions are expressed (i.e. whether they are reported as opinions held by others, or as the author's own opinions; cf. Busse and Schröder 2009), the nature of those opinions (i.e. whether they approve of problematic variants or not; cf. Albakry 2007), and the approach to using sources (i.e. whether pronouncements are based on sources or not; cf. Peters and Young 1997). Consequently, what these studies have also illustrated is that there is no single indicator of prescriptivism (cf. Peters and Young 1997), and no single way in which usage guides discuss usage, present opinions or facts, and offer advice. Building on these insights, the aspects I investigated in the usage guide entries are the following: the way usage guide authors treat the usage features, the attitudes expressed towards these features, and the dimensions of usage invoked in the treatment.

The analysis of the treatment of the six usage features in the collection of American usage guides was done by distinguishing three types of treatment: ACCEPTABLE, RESTRICTED, and UNACCEPTABLE.[2] This analysis of treatment is similar to the approach taken in previous studies of normative grammars and usage guides. Dekeyser (1975), for instance, establishes a number of methodological approaches in dealing with this issue empirically, which are reflected in some later studies. A case in point is his analysis of prescriptions found in normative grammars, using the following categories: $+$, if a grammarian supports a prescription, $-$, if a grammarian rejects a prescription to accept a problematic construction, or $\pm$, if the grammarian's opinion is between these two positions. A similar approach is used by Yáñez-Bouza (2015) for the analysis of normative grammar pronouncements on preposition stranding. This tripartite categorisation has also been used in a number of studies of usage guides.

The first extensive study to include such a classification is Creswell's (1975) analysis of the pronouncements found in the usage notes of the *American Heritage Dictionary* (1971). Motivated by the comparison of labelling practices in dictionaries by McDavid (1973), Creswell (1975) set out to investigate the extent to which dictionaries and usage guides agree in their usage judgements. He took the usage notes of the *American Heritage Dictionary* (1971) as a starting point for his comparison of the judgements on usage across a collection of handbooks of usage, and analysed them according to a number of dimensions. In the context of how various locutions were treated across different works, Creswell (1975) used the categories Not Treated ("either the word is not entered at all or, if it is, the specific problem in usage is not referred to"), Accepted ("either entered without comment, or discussed and approved, in the usage books the latter only"), and Restricted ("either assigned a restrictive label or or discussed and recommended to be completely avoided or limited to use in certain contexts") (Creswell 1975: 8). A similar approach was taken by Berk (1994) for the analysis of 26 books on usage. The distinctions in treatment she used are: "rule invoked", "rule rejected", "rule invoked for formal discourse", "rule may be overridden for rhetorical concerns", and "no discussion of rule" (Berk 1994: 111). In a study looking at the sources of evidence used in language advice literature in Australia, the United Kingdom, and the United States, Peters and Young (1997) used similar categories to those used by Creswell (1975), *viz.*, U for "unacceptable", A for "acceptable", and R for "usable in restricted contexts", where the final category

[2] I use small capitals for terms referring to analytical categories used in the present study. I use regular font when referring to the general notion of, for instance, acceptability.

"was applied whenever the usage writer explained some constraint on the usability of the structure", and was also used to "represent the rather equivocal stance of an author who admits that a certain usage practice is common but advises that 'careful writers do otherwise'" (Peters and Young 1997: 320–321). More recently, Albakry (2007) uses similar categories to analyse the treatment of five usage features in 18 usage and style guides in American English. Based on Peters and Young's categories, Albakry (2007: 34) distinguishes among "not acceptable (i.e. deemed to be incorrect and should be avoided), acceptable (i.e. deemed correct and should not be avoided), vague (i.e. commentator explains some constraints on the use of the structure or espouses an equivocal stance towards it) and not mentioned (i.e. the usage feature is not commented on in the particular usage guide)".

Given the discrepancies between these categories in previous studies, I provide more specific definitions of the three categories of treatment formulated for the purposes of this analysis, although they do not depart greatly from the basic distinctions. A treatment was classified as ACCEPTABLE when the author explicitly approves of the construction, as exemplified in (31) and (32). It is important to point out that cases in which some restrictions are mentioned, but where these restrictions are linguistic rather than social or situational, were also classified as ACCEPTABLE. Examples of this kind of entry are especially common in the treatment of the split infinitive, where the restrictions on the use of the feature have to do with the length of the element that separates the particle *to* from the verb, or with the naturalness or awkwardness of a construction, but not with whether it is socially or situationally appropriate. Such entries were classified as ACCEPTABLE, because they explicitly express acceptability of the feature across registers while sometimes also explicitly dismissing, and even disparaging, the prescriptive rule against it. The treatment was classified as RESTRICTED when the author partly approves of the construction, while noting restrictions on its use which are social or situational. This includes cases where the construction is criticised, but where it is also noted that the item is used in certain contexts, or when various opinions, both accepting and not accepting the feature, are mentioned, as in (33) and (34). Furthermore, entries that neither accept nor dismiss a feature were also put into the category RESTRICTED. This category may also be considered the most loosely defined, because RESTRICTED entries which do not contain any explicit judgement of a feature do not offer much evidence of attitudes, as in examples (35) and (36). Finally, entries in which the author explicitly disapproves of the construction and does not find it acceptable in any context were classified as UNACCEPTABLE; an example is given in (37).

(31) Do not be afraid to split infinitives or other verb forms. (Clark 2010: 255)

(32) I know of no usage authorities who believe that split infinitives are always wrong, but I take a more extreme position than most: More often than not, in my opinion, infinitives are better split. (Walsh 2004: 64)

(33) This word is a contraction of *am not* or *are not*, and can, therefore, be used only with the singular pronouns and *you*, and with the plural pronouns *we*, *you* and *they*, and with nouns in the plural. (Bechtel 1901: 119)

(34) Even though it has been universally condemned as the classic mistake in English, everyone uses it occasionally as part of a joking phrase or to convey down-to-earth quality. But if you always use it instead of the more "proper" contractions you're sure to be branded as uneducated. (Brians 2003: 6)

(35) *Literally* means "actually, without deviating from the facts," but it is so often used to support metaphors that its literal meaning may be reversed. In statements like the following, *literally* means "figuratively" and *literal* means "figurative":

The Village in the twenties [was] a literal hotbed of political, artistic, and sexual radicalism.–Louise Bernikow, New York Times Book Review

In this struggle, women's bodies became a literal battleground.–Martin Duberman, ibid.

[New York City is] literally hanging by its fingernails.–Walter Cronkite, CBS News

Literal-minded readers find such locutions absurd. (Ebbitt and Ebbitt 1978: 547–548)

(36) Writers are so often besought by rhetoricians not to say *literally* when what they mean is *figuratively* that one would expect them to desist in sheer weariness of listening to the injunction. The truth is that writers do not listen; and *literally* continues to be seen as a mere intensive that means *practically*, *almost*, *all but*. *He was literally speechless. He could only murmur: "Good God!"* This speechlessness would be literal only if he had been incapable of uttering the words we are told he murmured. [A golf cart] *literally floats over the roughest fairway*. To accomplish this it would have to be one of those vehicles that ride a few inches above the ground on a cushion of air. Since this particular cart moves with its wheels on the ground, the floating is figurative. (Follett 1966: 204)

(37) This cannot be called a contraction, and however much it may be employed it will still be only vulgarism. *I'm not* is the only possible contraction of *I am not*, and *we're not* of *we are not*. (Ayres 1911: 6)

The second dimension analysed was the attitudes to the language features expressed in the usage guide entries. The general notion of attitudes notoriously defies a straightforward definition, so it is important to distinguish between attitudes to usage in usage guides and speakers' attitudes in the context of this analysis. The issue of the multifarious nature of attitude studies was referred to in Section 2.7 above. At the end of that section, I referred to a group of studies in which the term "attitudes to usage" has been used in relation to the attitudes of grammarians, writers on language, or language authorities (cf. e.g. Finegan 1980; Sundby et al. 1991). In this sense, then, the notion of "attitudes to usage" refers to attitudes expressed in metalinguistic publications, such as grammar books, style guides, language manuals, and usage guides, and can be understood as an instantiation of metalinguistic commentary; as such, this notion of attitudes to usage should be distinguished from speakers' attitudes. While this is indeed an important distinction, the term "attitudes" seems appropriate in the context of my analysis of usage guides, because it captures the subjective and attitudinal component of the pronouncements found in usage guides.

This analysis is based on an approach which has been employed in studies of attitudes to language in normative grammars (cf. e.g. Sundby et al. 1991), and which provides a useful point of departure for the analysis of attitudes in usage guides. Important in this respect are analyses of normative or prescriptivist metalanguage, that is, the various ways in which grammarians expressed their ideas about language. Studies on normative grammar in eighteenth-century English that are important in the context of attitude analysis are mainly those that deal specifically with the labels used by grammarians in the treatment of linguistic variants. Sundby et al.'s *Dictionary of English Normative Grammar* (*DENG*) (1991) marks an important step in providing a fairly comprehensive inventory of the "prescriptive labels" that comprised the metalinguistic system of eighteenth-century grammarians. The classification presented in *DENG* provides the opportunity to use those labels to study attitudes to usage. A good example of this is Yáñez-Bouza's (2015) study of the attitudes of normative grammarians towards preposition stranding in the eighteenth century. By relying on labels used in grammars, she shows how labels found in normative grammars reveal the emergence of strikingly conservative attitudes towards preposition stranding that arose in the middle of the eighteenth century. Building on Sundby et al. (1991), Yáñez-Bouza (2015) provides a comprehensive list of terms that expressed attitudes to preposition stranding in that period, and classifies them according to whether they can be interpreted as advocating or criticising the construction. The analysis of such labels is thus part of the present approach.

The categories used here were POSITIVE and NEGATIVE attitudes, exemplified in (38) and (39) below. These two categories were taken as the most intuitive way of analysing the attitudes expressed towards the features. Attitudes to usage in usage guides are usually expressed through a kind of semi-technical language consisting of a number of metalinguistic expressions such as "vulgarism", "error", "gross linguistic gaffe", "natural", "acceptable", and so forth. More rarely, attitudes can also be expressed by the use of particular verbs, such as "avoid". In all instances of explicitly expressed attitudes, the distinction between POSITIVE and NEGATIVE attitudes is fairly intuitive, and was taken to serve as a basis for the present analysis.

(38) Despite the taint of *ain't* from its origin in regional and lower-class English, and more than a century of vilification by schoolteachers, today the word is going strong. It's not that *ain't* is used as a standard contraction for negated forms of *be*, *have*, and *do*; no writer is that oblivious. But it does have some widely established places. One is in the lyrics of popular songs, where it is a crisp and euphonious substitute for the strident and bisyllabic *isn't*, *hasn't*, and *doesn't*, as in "It Ain't Necessarily So," "Ain't She Sweet," and "It Don't Mean a Thing (If It Ain't Got That Swing)." (Pinker 2014: 204)

(39) *Ain't* is unrecognized in modern English & as used for *isn't* is an uneducated blunder and serves no useful purpose. (Nicholson 1957: 49)

The third aspect of the analysis is related to the dimensions of usage identified by usage guide writers. In his account of writing a usage guide, Weiner (1988) uses the term "sociolinguistic considerations" to refer to extralinguistic factors that usage guide writers draw on in their selection and discussion of usage problems, as well as the information provided about those extralinguistic factors in treatments of usage. It is, of course, important to note that the term "sociolinguistic considerations" does not imply that such observations made in usage guides are to be understood as objective, or as based on descriptive or empirical studies. Nevertheless, these observations are important in showing how certain sociolinguistic aspects of the features are referred to and understood by usage guide writers (cf. Tieken-Boon van Ostade 2015; Kostadinova 2018a). What Weiner (1988) refers to as "sociolinguistic considerations" can be compared to levels of usage or usage dimensions, a question usually discussed in the context of dictionaries. In dictionaries such considerations tend to be expressed through a set of labels, such as "common", "rare", "archaic", "dialectal" (cf. McDavid 1973; Creswell 1975; Card et al. 1984). Sundby et al. (1991: 38) similarly refer to the

practice of eighteenth-century grammarians of referring to a number of dimensions in their treatment of usage, such as "medium ('we never write'), genre ('hardly allowable in poetry'), frequency ('seldom used'), attitude ('rude especially to our betters'), social position ('low'), linguistic competence ('adopted by the ignorant'), territory ('peculiar to Scotland'), etc.".

These kinds of dimensions of usage are also characteristic of usage guides, and were consequently analysed separately. Employing both of these terms in the discussion of the treatment of linguistic features in usage guides presents certain problems. The term "sociolinguistic considerations" may imply a scientific sociolinguistic basis for those considerations, which, in reality, is both hard to prove and very unlikely. Even in cases where usage guide writers have attempted to consider sociolinguistic aspects of the use of particular features, sources are rarely cited (Peters 2006; Peters and Young 1997), which makes it difficult to rely on the unproven presupposition that any kind of reference to sociolinguistic factors should be understood as referring to actual sociolinguistic processes or phenomena. The problem with dimensions of usage used in dictionaries is that they tend to be fairly well formalised and strictly defined, usually in the context of one particular dictionary. In general, however, dimensions of usage are expressed through different labels in different works and in different time periods. There is thus often disagreement as to what the specific levels of usage should be. Given all of these considerations, the term "dimensions of usage" seems more appropriate than "sociolinguistic considerations", and is therefore used in the present discussion. The dimensions of usage in the entries were analysed by annotating the entries for one of the following six categories: FREQUENCY, MODE, REGISTER, SPEAKERS, VALUE, and VARIETY. Examples of these are given in (40)–(45), respectively.

The tag FREQUENCY was given to statements about the use and the frequency of use of a feature, exemplified in (40). MODE refers to mode of expression, or reference to writing or speech, or both (41). A similar category is used in Creswell's (1975: 24–26) analysis of the usage notes in the *American Heritage Dictionary*, where he found that in the usage features on which the *AHD* panel was asked to vote, in more than half of the cases it was not specified whether the questions of usage on which the panel voted referred to speech or to writing. This is perhaps indicative of the lack of inclusion of levels of usage in usage discussions and usage advice, which may in turn produce misunderstanding about the use of a particular feature, and thus make the interpretation of the votes of the panel problematic. The presence of this aspect in usage guides may thus be seen as an important refinement in the treatment

of language usage. REGISTER was used to label statements about contextual or situational aspects of using a certain feature, as shown in (42). The tag SPEAKERS is used to describe references to groups of people identified in the entries (43). VALUE is used to refer to a wider more general meaning of social value associated with the use of a feature. The example in (44) illustrates this dimension. Finally, VARIETY is used to classify all references to a specific regional or social variety of English (45).

(40) Today, split infinitives continue to appear often in Standard speech and even in Edited English. (Wilson 1993: 22)

(41) These uses of *like* are typical of informal spoken language, especially of younger people, and their occurrence in writing is limited chiefly to dialogue. (Pickett et al. 2005: 282)

(42) Some authorities feel that *ain't* would be a useful addition to informal English, particularly as a contraction for *am I not*, which has none that can be pronounced easily. (O'Conner 1998: 128)

(43) Its use is pretty much confined to users of standard English and to literary contexts. (Gilman 1989: 867–868)

(44) Like parallel fifths in harmony, the split infinitive is the one fault that everybody has heard about and makes a great virtue of avoiding and reproving in others. (Follett 1966: 313)

(45) Standard use is hard to explain, but clearly Americans have come down hardest on it, and they have made the rejection stick in Standard American English. (Wilson 1993: 22)

The entries were annotated using the annotation tool 'brat' (Stenetorp et al. 2012),[3] based on the three levels of analysis explained in detail above. This multi-level annotation allows for specific kinds of information to be extracted from the usage guides and to be analysed side by side; these kinds of information include, but are not limited to, the ways in which levels of usage are conceptualised in usage guides, and which groups of speakers or varieties of English are the ones most often referred to. Considerations of register and mode of communication may reveal a less conservative account of usage, while considerations of social value present in guides serve as an important basis for a comparison of these values with attitudes held by speakers. The results of such an analysis, as I will illustrate in the next chapter, show that usage

[3] Available online at http://brat.nlplab.org/.

guide writers, regardless of their tone, often draw on various social aspects of the use of the features they discuss. Even though these considerations may be subjective, they also serve as important indicators of more general opinions about how features are used. Finally, the inclusion of various types of references to dimensions of usage will be shown to be a crucial aspect of the development of the genre towards presenting a more varied and nuanced account of language use.

4.4 Actual language use: data and analysis

My approach to the analysis of the influence of prescriptivism on language change conceptualises language change by drawing on descriptive or usage-based accounts of language change, which analyse frequency patterns of linguistic features. Patterns of variation and change are thus determined by correlating linguistic variables with extralinguistic factors such as time, genre, register, etc. (Nevalainen 2006a: 560–561). Studies of language variation and change in this vein, associated for instance with historical sociolinguistics, have established that when it comes to language change, "the picture that emerges is one of gradual evolution rather than abrupt change" (Mair and Leech 2006: 319). Language variation and change is thus assessed on the basis of identifying "shifting frequencies of use for competing variants", one of which is prescriptively targeted (Mair and Leech 2006: 319); this allows for the importance of prescriptive influence or effect to be assessed. The approach taken in sociolinguistic and descriptive corpus-based research on language variation takes changes in frequency patterns to be indicators of language change (cf. Mair 2006: 2). Consequently, if we take this approach to language change, and if we manage to show that changes in frequencies across time are constrained by prescriptivism-related factors, we can posit an influence of prescriptivism on language change.

The data for the analysis of the actual usage patterns for each feature were extracted from the Corpus of Contemporary American English (COCA, Davies 2008–present) and the Corpus of Historical American English (COHA, Davies 2010b), both created and maintained by Mark Davies at Brigham Young University. Although both corpora are available for online use via the BYU interface, the analysis conducted for this study required the use of the full-text data, which can be purchased in three different formats: plain text, word/lemma/part-of-speech-tagged files, and SQL-database format. This makes wide-ranging exploration of the data possible, in terms of the analysis of many different features, as well as many different aspects

of the texts included in the corpora, such as the type-token ratio of each separate text in the corpus. The full-text data come with some limitations, due to texts in the corpora being affected by copyright restrictions. Consequently, the version of the corpus data available for purchase has been transformed, in order to abide by copyright restrictions. This transformation consists in the removal of ten out of every 200 words of text. This ensures legal use of the copyrighted text in such a way that the strength of the corpus and the validity of the results obtained are not seriously affected, because, as stated on the website, "95% of the data is still there".[4] This means however, that the results obtained in this analysis will be somewhat different, though not significantly so, from the same results obtained through the corpora's online interface.

The COCA corpus is described by its creator as "the first reliable monitor corpus of English" spanning the period 1990–2015 (Davies 2010a: 447; see also Davies 2008–present, 2009), and, as a monitor corpus, it is regularly updated with new data. As of March 2017, according to its website, COCA contained more than 520 million words of text, or 20 million words for each year. The full-text corpus data used for this analysis cover the period 1990–2012; thus the total number of words is lower than 520 million. The COCA corpus is divided into five sections: academic, fiction, magazines, newspapers, and spoken. Davies (2009: 161–162) describes in more detail the selection and sources of materials in COCA, and the specific subcategories for each section are given in Table 4.2. COHA can perhaps be described to some extent as the historical counterpart to COCA. It contains 400 million words, divided into four sections: fiction, magazines, newspapers, and non-fiction books (Davies 2012). These four sections are further subdivided into subgenre categories, given in Table 4.2. A complete list of sources for both corpora is available on the website, while Davies (2009, 2010a, 2012) describes in detail how the corpora have been created and how they can be used to study language variation and change in different ways.[5]

It is important at this point to raise the question of what the different corpus sections, and their subcategories, represent in terms of register variation in historical and present-day American English. The understanding of the nature of texts in the two corpora is instrumental in interpreting the results, for two main reasons. The first one relates to the comparability of the two corpora. In other words, it is important to understand in what ways the two corpora are similar, as well as where they differ, so that the interpretation of the results can take this into account. The general intention

[4]For more details see http://corpus.byu.edu/full-text/limitations.asp.

[5]I use the terms 'genre' and 'register' interchangeably when referring to the various corpus sections.

COCA	*COHA*
Academic	**Non-fiction**
History	
Education	
Geography/Social Science	
Law/Political Science	
Humanities	
Philosophy/Religion	
Science/Technology	
Medicine	
Miscellaneous	
Fiction	**Fiction**
General (Books)	Drama (Plays)
General (Journal)	Movie Scripts
Science Fiction	Novels
Juvenile	Poetry
Movie Scripts	Short stories
Magazine	**Magazine**
News/Opinion	No subgenres
Financial	
Science/Technology	
Society/Arts	
Religion	
Sports	
Entertain	
Home/Health	
African-American	
Children	
Women/Men	
Newspaper	**Newspaper**
Miscellaneous	No subgenres
International News	
National News	
Local News	
Money	
Life	
Sports	
Editorial	
Spoken	
No subgenres	

Table 4.2: Sections and subsections of COCA and COHA

underlying the use of these two corpora was that they would provide comparable evidence, at least for some registers, so that the development of usage could be traced over a longer period of time. In the case of fiction, for instance, this is quite straightforward (Table 4.2).

In terms of comparability, COCA and COHA contain similar material for the fiction, magazines, and newspapers sections in both corpora. For the historical corpus, some text types, such as movie scripts, newspapers, and magazines, are only available from the twentieth century onwards. A spoken section is only available in COCA, so my analysis of spoken data is restricted to the period 1990–2012. Finally, COCA has an academic section, while COHA has a non-fiction section. While these two are different categories, it seems a priori not unreasonable to assume that they are not dissimilar, given that they both contain a preponderance of relatively formal texts. As we will see in Chapter 6, this assumption is supported by the frequency patterns observed in these two genres. An additional reason for this consideration is that the organisation of the subcategories of texts for both the academic section of COCA and the non-fiction section of COHA is according to the general Library of Congress classification system (cf. Davies 2009: 162; Davies 2012: 124–125).

Finally, a limitation of COCA, when it comes to working with these subcategories, is that the subsections included in the spoken section are not based on actual genres of spoken language, but on the sources from where the texts were obtained. On the naturalness of the spoken data, Davies notes that they "are based almost entirely on transcripts of *unscripted* conversation on television and radio programs" (Davies 2009: 162) and that these texts are accurate and spontaneous; in other words, they provide a good representation of non-media English. The corpus data were then used to extract all occurrences of the various linguistic features with the help of Python scripts. In most cases, occurrences were extracted from the part-of-speech-tagged files of the corpora, using regular expressions. The use of the part-of-speech-tagged data greatly facilitated further analysis of each occurrence.

The data analysis can be divided into two parts. In the first part of the analysis, I look at the frequency of occurrence of a particular feature across subsections of the corpora, as well as across time periods. For the analysis of patterns of use, I will rely on the two approaches used in corpus-based linguistic analysis of variation and change suggested by Biber et al. (2016): the text-linguistic approach and the variationist approach. As Biber et al. (2016) show, these two approaches can sometimes produce different results, as an analysis of the frequency distributions of one variant is a good indication of the rate of occurrence of that variation, while the results of a

variationist analysis allow us to "report proportional preference; they do not actually tell us how often a listener/reader will encounter these structures in texts" (Biber et al. 2016: 359). Leech and Smith (2009: 176–178) also discuss the difference between variationist and text-linguistic measures of frequencies of usage, but use the terms "proportionate" method and "normalisation" method to refer to variationist and text-linguistic approaches, respectively. What a variationist account shows us is how often a variant occurs, in the context of two or more options, out of a total number of possible occurrences. In the second part of the analysis, I specifically look at whether and how additional prescriptively targeted features may predict the use of a proscribed feature. This analysis is done using the split infinitive as a case study. In what follows, I discuss each of these two types of analysis in more detail.

When establishing the variable context, it is important to distinguish between contexts in which variation is a choice, as opposed to contexts where one variant is categorically used (cf. Poplack and Dion 2009: 571: "in contexts where speakers must choose among the major variants"). With reference to the features used in this study, the assumption is made that theoretically, both the standard and the proscribed variants are possible options, given that the notion of 'incorrectness', prescriptively speaking, is linguistically arbitrary. Of course, establishing variants was not possible in the context of the discourse particle *like* and non-literal *literally* (cf. Section 2.6), since it is difficult to conceptualise a variable in which *like* is one of the variants; in addition the environments in which this variant could occur would be almost impossible to predict or determine (but see D'Arcy 2007).

The analysis of *ain't* was done on the basis of both text-linguistic and variationist frequencies. These were analysed in the context of two variables: present tense negative *be not* and *have not*. The variants are given in Table 4.3. It is also important to address a point of disagreement present in the methodological approaches employed in previous studies to determine the variables in which *ain't* is used, that is, the forms with which it alternates. Wolfram (1974: 153), for instance, considers *ain't* as alternating only with the *'m* / *'re* / *'s not* forms of *be not*, noting that *aren't* and *isn't* were not observed in his sample of speech from Puerto Rican English speakers, and were thus excluded from the analysis. Cheshire (1981) considers *ain't* as a variant of all contracted forms of present *be not* and *have not*, including both auxiliary and copular *be not*, but not the full forms of these verbs (this is not stated explicitly, but can be inferred from the data presented). Weldon (1994) considers *ain't* as a variant of both full and contracted copula in *be not* environments, while Anderwald analyses the alternations of *ain't* with *be not* and *have not*, including both auxiliary

and copula *be not*, but does not specify whether the total number of environments includes full forms. Taking into account these differing approaches, in my analysis I take both full forms and contracted forms of *be not* and *have not*, primarily because my data come from a general standard American English corpus, in which both full and contracted forms occur. Second, I do not distinguish between copula and auxiliary *be not* environments, because (a) such a distinction is not particularly relevant for the issue of the influence of prescriptivism on usage, and (b) studies of variation have shown that it is hard to identify regularities in the patterns of variation between copular *be not* and auxiliary *be not*, as opposed to *have not*, leading Anderwald (2002: 139) to conclude that "a distinction of BE into auxiliary and copular uses is perhaps not particularly warranted".

	be not		***have not***	
prescribed	*am / are / is not* *'m / 're / 's not* *aren't / isn't*	*I am not coming.* *I'm not coming.* *He isn't coming.*	*have / has not* *'ve / 's not* *haven't / hasn't*	*I have not left.* *I've not left.* *I haven't left.*
proscribed	*ain't*	*I ain't coming.*	*ain't*	*I ain't left.*

Table 4.3: The variants of *be not* and *have not*

For the analysis of the newer uses of *like*, as discussed in Section 3.4 above, I will focus only on the discourse particle *like*. While some of the occurrences of the discourse particle *like* are tagged with appropriate part-of-speech tags in the corpora used, it was noticed that the accuracy of the tagging was not very high, and that in most cases a much better indicator of whether *like* is used as a discourse particle or not was the transcription of the data. Accordingly, since the discourse particle *like* is set off with commas in the majority of the cases, this was used as a more reliable way of identifying and extracting those instances of *like*. In the cases of *like* and *literally*, for instance, a variationist analysis was not undertaken, due to the complicated issue of establishing the variable context (but see D'Arcy 2007 for a variationist analysis of vernacular *like*). In the case of *like*, only normalised frequencies of occurrence were therefore used to track the patterns of usage of this feature, based on the tags in the corpora.

In the case of *literally*, two analyses were carried out: one on the basis of the entire set of occurrences of *literally*, and another on the basis of a subset of occurrences. This decision was made in view of the fact that the process of change which *literally* is undergoing entails a significant amount of variation in its uses, and, consequently, it

is difficult to neatly disambiguate the meanings and functions of these uses, especially of the newer ones. The three uses of *literally* distinguished in the analysis are given in Table 4.4; these were distinguished on the basis of categorisations of the uses of *literally* found in previous studies (see Section 3.5).

	uses of *literally*	
prescribed	primary	*This is literally translated.*
proscribed	dual	*There were literally thousands of people.*
	non-literal	*This book literally blew my mind.*

Table 4.4: The uses of *literally*

In its primary use, *literally* refers to what something means, how something is said or meant, or how something is translated or interpreted. In such cases, *literally* has a clear denotational meaning, as it functions as a manner adverb; here *literally* is the answer to the question how something is done (e.g. *How do you mean? I mean literally*.) In all other cases, however, the meaning of *literally* is more ambiguous and elusive, and in such cases it is almost always possible that the speaker is using *literally* to signal that something that may be understood figuratively must be understood literally (cf. creative cases in Powell 1992, discussed in Section 3.5 above). In these cases, the meaning of *literally* is highly dependent on context. In view of this, a manual analysis of all occurrences of *literally* was deemed too laborious to be worthwhile, given that it would be unlikely to reveal any major insights. A middle-ground solution was to perform both an automatic binary disambiguation of the entire set of occurrences of *literally* using Python scripts and a manual analysis classifying a subset of the occurrences of *literally* into three categories. The first type of analysis distinguishes between cases that are very clearly instances of the so-called primary use of *literally*, and all other cases (for more details on how this was done, see Appendix C). An important consideration in this decision was the fact that prescriptive attitudes are highly conservative with respect to most types of changes in the language. In the treatment of *literally*, especially, only the primary (i.e. denotational) meaning of the word was therefore accepted, while all other uses (i.e. intensifying, subjective or metapragmatic ones) are seen as 'incorrect'. This first part of the corpus analysis is based on data from both corpora, and allows us to track the usage of the proscribed variants over time.

The manual analysis was carried out on a sample of the total number of occurrences of *literally* in the corpora. This sample was extracted by selecting every

fifth case in the set of all occurrences of *literally* in both COCA and COHA. This manual analysis was done for two reasons. First, it was meant to provide additional evidence for the distribution of the three uses of *literally*, which was not possible with the automatic disambiguation of all its occurrences. Second, the manual analysis also served to identify any possible differences between the results of the automatic disambiguation and the manual analysis of the data. In this way, I hope to be able to provide a rough estimate of the level of accuracy of the automatic disambiguation.

An analysis of the frequency patterns of negative concord also requires a variationist account of the ratio of the stigmatised forms in comparison to all others; here, we cannot rely on frequency counts of negative concord constructions alone. In order to delimit the total number of possible environments in which negative concord may occur, it is important to consider the variants of negative expression and the circumstances in which negative concord can be expected. The rule of negative attraction (first formulated by Klima 1964: 267, 289, cited in Labov et al. 1968: 268; see also Wolfram 1974: 163) is the starting point for determining the contexts for negative concord. According to this rule, in standard English sentences with indefinites, there are two possible options. If the indefinite is in pre-verbal position, the negative marker is attracted to the first indefinite before the verb, which accounts for sentences such as *Nobody knows anything*, to use Labov et al.'s example (1968: 268). If the indefinite is post-verbally located, the negative marker may optionally be attached to the verb, as in *John doesn't know anything*, or to the indefinite, as in *John knows nothing*. In other words, if the indefinite comes after the verb, "the negative attraction rule may or may not apply" (Wolfram 1974: 164). In relation to these two variants, Wolfram (1974: 165) further notes that the latter is more characteristic of literary than of colloquial English. In African American English, as well as in non-standard varieties of English, the rule of negative attraction does not apply in cases where post-verbal indefinites keep the negative marker, resulting in instances such as *John doesn't know nothing*. In this case, "what takes place is a copying of the negative on as many post-verbal negatives as there are in a sentence" (Wolfram 1974: 165). In applying this rule to determining potential environments in which negative concord can occur, so that we can establish the ratio of usage of negative concord (cf. Smith 2001; Nevalainen 2006b), we can employ the variants given in Table 4.5. Since my goal here is not to give an account of negative concord in the entire language system, in order to simplify the analysis, I limited myself to investigating negative concord occurrences in sentences with the following indefinites: *anybody/nobody*, *anyone/no one*, and *anything/nothing*.

	V + INDEFINITE	
prescribed	V-neg + INDEFINITE	*I haven't seen anybody.*
	V + INDEFINITE-neg	*I have seen nobody.*
proscribed	V-neg + INDEFINITE-neg	*I haven't seen nobody.*

Table 4.5: The variants of negative concord

Pronouns in coordinated phrases were extracted only with first person pronouns. The contexts of variation were further limited to include only certain types of cases in which pronouns are used in coordinated phrases. For instance, the analysed cases of object *I* and subject *me* are only those cases in which the pronouns *I* and *me* are found in coordinated phrases where the other phrase-constituent is a proper noun. This decision was made in light of the fact that there may be additional constraints affecting the realisation of *I* or *me*, especially if the other constituent is another pronoun. Phrases with proper nouns were seen as presenting a sufficiently uniform context in which the realisation of *I* or *me* will not be expected to be affected by the case of the other phrase-constituent. The secondary reason for this decision was of a practical nature. The identification of coordinated phrases functioning as subjects or objects in which one of the constituents in the phrase is *I* or *me* was not a straightforward task of automatic extraction from the corpora. The restriction to cases with proper nouns significantly reduced the danger of extracting a large number of false positives from the data, without influencing the quality of the data. The variants for pronouns are given in Table 4.6.[6]

	Subject		**Object**	
prescribed	*x and I*	*Elly and I left.*	*x and me*	*They saw Elly and me.*
	I and x	*I and Elly left.*	*me and x*	*They saw me and Elly.*
proscribed	*x and me*	*Elly and me left.*	*x and I*	*They saw Elly and I.*
	me and x	*Me and Elly left.*	*I and x*	*They saw I and Elly.*

Table 4.6: The variants of first person pronouns in coordinated phrases

The split infinitive was also analysed on the basis of both text-linguistic and

[6]The table contains *I and x* as a 'prescribed' variant simply because the form of the first person pronoun adheres to the prescription that the nominative form should be used in subject positions. However, it should be noted that there is a different set of norms against this use, mostly having to do with the impoliteness of referring to oneself first. I have not taken this into account in the analysis; however, I do not consider this a significant influence on the results, as this variant was extremely rare in the corpus data.

variationist frequencies. To establish the text-linguistic frequencies, all infinitives split by one modifier were extracted from the data. The variationist frequencies were established only on the basis of a specific variable context, which was established as infinitives modified by one lexical adverb ending in *-ly*. The reason for this was that different elements which may be placed between the *to* and the infinitive behave differently. Good examples of this are cases in which *not* is placed between *to* and the infinitive verb. In these cases, there are two possible variants: *not* + *to* + verb or *to* + *not* + verb, which is a pattern of variation different from infinitives modified by a *-ly* adverb, because *-ly* adverbs can be placed after the verb, in addition to before the verb and before *to*. The variants thus established are given in Table 4.7.

	MODIFIED INFINITIVE	
prescribed	*to* + INFINITIVE + modifier	*to improve significantly*
	modifier + *to* + INFINITIVE	*significantly to improve*
proscribed	*to* + modifier + INFINITIVE	*to significantly improve*

Table 4.7: The variants of the modified infinitive

On the basis of the variables outlined here, the occurrences of these linguistic features were extracted from the corpus data. The sizes of the various datasets differed; Table 4.8 gives an overview of the total number of occurrences of each feature in the corpus data. More specific information on sample sizes, as well as the number of occurrences of prescribed, as opposed to proscribed, variants are given in the relevant section in Chapter 6.

Feature	**COHA**	**COCA**
ain't	39,348	12,228
like	634	10,020
literally	6,848	14,946
negative concord	10,041	8,530
object *I*	194	380
subject *me*	456	819
split infinitive	10,062	63,079

Table 4.8: Raw frequencies for each of the features extracted from COHA and COCA

The second part of the corpus analysis which I mentioned above focuses specifically on investigating the relationship between proscribed variants of a particular feature and the use of other prescriptively targeted features. This analysis

is based only on the COCA data, and it is conducted using the split infinitive as a case study. The general approach is based on Hinrichs et al. (2015), who distinguish between intralinguistic and extralinguistic constraints on the variation in the realisation of the relative pronoun in restrictive relative clauses. In addition to the intralinguistic and extralinguistic constraints, Hinrichs et al. (2015) also included a number of prescriptivism-related constraints, expressed as frequency of occurrence of other prescriptively targeted features. This was done on the basis of the assumption that if a feature's variation is affected by prescriptive constraints, then these constraints would also be noticeable in the context of other features that might be expected to be affected by prescriptivism. They thus included four features in their multivariate analysis. In the present analysis, the same principle was used; however, the number of additional prescriptivism-related predictors was increased. The goal of this analysis is to investigate the potential influence of prescriptivism at the level of individual texts and on the basis of a number of different language features. In this way, it is hoped that this analysis will supplement the separate analyses on the six language features which this study focuses on.

The dataset used for this analysis was composed of all the occurrences of infinitives modified by a single *-ly* adverb extracted from COCA. The dependent variable was defined as MODIFIED INFINITIVE, with two levels: SPLIT and NON-SPLIT. Thus, each occurrence of a modified infinitive in the dataset was classified at either of the two levels. A set of additional predictors were defined, as outlined in Table 4.9; each occurrence of a modified infinitive in the dataset was additionally coded for each of these predictors. In terms of internal predictors, I distinguish the semantic class of the adverb and the length of the adverb compared to that of the verb, i.e. the independent variables ADVERB TYPE and ADVERB LENGTH, respectively. The ADVERB TYPE for each occurrence of a modified infinitive was determined on the basis of the semantic classification of adverbs in Biber et al. (1999: 552–560). The adverbs modifying the infinitive were classified into the following categories: ADDITIVE-RESTRICTIVE adverbs (e.g. *especially*), DEGREE adverbs (e.g. *almost*), LINKING adverbs (e.g. *therefore*), MANNER adverbs (e.g. *happily*), STANCE adverbs (e.g. *probably*), and TIME adverbs (e.g. *recently*). The length of the adverb was operationalised as the number of syllables and as a categorical variable with three levels: SHORTER, if the adverb is shorter than the verb measured in number of syllables; EQUAL, if the adverb has the same number of syllables as the verb; and LONGER, if the adverb has more syllables than the verb. External predictors included in the analysis are YEAR and GENRE. YEAR is a continuous variable with

values from 1990–2012, associated with the source year of each corpus text in which relevant instances of modified infinitives were identified. The predictor GENRE is a categorical variable with five levels, corresponding to the five corpus sections: ACADEMIC, FICTION, MAGAZINE, NEWSPAPER, SPOKEN.

Predictors	**Levels**
Internal predictors	
ADVERB TYPE	ADDITIVE-RESTRICTIVE DEGREE LINKING MANNER STANCE TIME
ADVERB LENGTH	LONGER EQUAL SHORTER
External predictors	
YEAR	1990–2012
GENRE	ACADEMIC FICTION MAGAZINE NEWSPAPER SPOKEN
Prescriptivism-related predictors	
And/But *data is* *hopefully* *less* + plural nouns *these kind of/these sort of* *none are* passives *shall* *try and* *whom*	frequency per 1,000 words

Table 4.9: Predictors used in the analysis of prescriptive constraints on the use of split infinitives

Finally, prescriptivism-related predictors are a number of additional features that are also often proscribed in usage guides. The additional features were selected on the basis of their frequency of occurrence in the prescriptive literature, as well as their relative ease of analysis. The HUGE database, described in more detail in the previous section, was the tool used to assess the most commonly treated features in

usage guides. On the basis of a search in the HUGE database, a list of features can be extracted and ordered by the number of guides in which the particular features are treated. Features were then further selected in view of their relative ease for automatic disambiguation. The additional features thus selected are: the use of *and/but* at the beginning of the sentence, singular *data*, *hopefully*, *less* with plural nouns, *these kind/sort of*, *try and*, plural *none*, passives, *shall*, and, finally, *whom*. In addition to these features, *ain't*, *like*, *literally*, and negative concord were also included in the dataset. For each text in the corpus data, the frequency of each of these features was established; the raw frequencies were normalised per 1,000 words. In order to obtain a more uniform dataset, texts from the corpus whose total number of tokens was too low (e.g. 500 words) or too high (e.g. 90,000 words) were excluded from this analysis, because the frequency of occurrence of the features used as predictors would be affected by the differences in size of the texts, especially when it comes to short texts, as this produced many zeros in the dataset. The final dataset contained 4,925 occurrences of a modified infinitive across the same number of texts. The analysis applied to this dataset was fixed-effects binomial logistic regression. Logistic regression is a technique applied in cases where the outcome variable, or the dependent variable is a one of two possible values – in this case SPLIT or NON-SPLIT infinitive (see Baayen 2008: 195). The technique is used to estimate the probability of one of these two possible values in comparison with the other, given the set of predictors – in this case all the predictors described above. In other words, the technique allows us to investigate questions such as: do texts in which, for instance, the frequency of *ain't* is high predict the probability of a modified infinitive being realised as split as opposed to non-split? The assumption for conducting such an analysis is that the observations are independent from each other. In order to satisfy this assumption, and make sure the cases are independent, only one case of a modified infinitive from each text was used.

4.5 Attitude survey: data and analysis

Finally, with a focus on present-day English specifically, data on the attitudes of speakers of American English towards the six linguistic features investigated in this study were collected and analysed. The data on speakers' attitudes were collected during a two-month fieldwork stay in Los Angeles, in 2014. The choice of Los Angeles was determined partly by the fact that it is one of the biggest metropolitan areas in

the United States, where I expected to find a great deal of variation in speech, and speakers, and partly because of contacts I had established there, which allowed me to recruit respondents more easily, and to enter communities as a 'friend-of-a-friend'. The main goal was to collect data from speakers who do not belong to the category of 'language professionals' and are not solely university students, as is the case with some previous studies on attitudes to usage (cf. Leonard 1932; Mittins et al. 1970; Albanyan and Preston 1998). Through a friend who worked at a start-up company in Beverly Hills, I interviewed a number of young adult professionals, who were chosen for my research in the attempt to collect data from respondents other than university students as well. Through my friend's family members, I interviewed a number of older professionals, and an additional number of respondents were recruited through the contacts that were established via these interviews. The aim was to arrive at a sample which is varied in terms of age. An additional number of respondents also came from Santa Monica College, where I distributed flyers to recruit potential respondents. Here, the focus was on recruiting first-generation students. The respondents received ten dollars for their participation, which lasted between 30 and 60 minutes. Table 4.10 shows the make-up of the sample of respondents. One limitation of the sample in this study is that it does not form a representative random sample of the population of Los Angeles. Not only was a fully random stratified sample beyond the scope of this data collection process because of time constraints, it would have also entailed the determination of categories of speakers a priori, an approach which raises its own methodological issues. In other words, such a sample would have meant that certain social categories or social variables would have needed to be defined in advance

The respondents were selected using a 'friend-of-a-friend' technique, resulting in a convenience sample of 79 respondents in total; their responses were used in the analysis of attitudes in Chapter 7. Table 4.10 shows that the sample of respondents is skewed towards younger adults. This limitation of the sample of respondents makes it difficult to carry out a comparison across all the age categories, but it does provide insights into how the attitudes of these speakers differ across usage features and contexts of use. The age categories presented in the table break down the sample by 10-year groups, but these were not used in the analysis of the data, where I divided informants into two age groups (see below, and Chapter 7).

Age group	Female	Male	African American	Other	White	Total
19–29	22	24	14	18	14	46
30–39	7	11	0	8	10	18
40 +	7	8	3	3	9	15
Total	36	43	17	29	33	79

Table 4.10: Distribution of respondents according to age, gender, and ethnicity

The meeting with each respondent consisted of two parts. The first part was a survey in which the respondent was asked to rate sentences containing the usage problems investigated in this study. The sentences used in this part of the survey were selected from COCA, described in the previous section of this chapter, and were used in different contexts, based on the corpus section in which they were found. This was done to ensure naturalness of the stimuli. Rather than presenting the respondent with a written choice of register (e.g. 'acceptable in formal writing', etc.), which is the approach taken in previous studies on attitudes to usage (cf. Mittins et al. 1970; Ebner 2017), the variable context of use was included as part of the stimulus. Thus, there were different types of sentence stimuli for each of the linguistic features investigated, in different contexts: 'spoken informal', 'spoken formal', 'written informal', and 'written formal'; the contexts for each of the sentences are given in Table 4.11. As the table shows, not all language features were included in the survey in all four contexts, simply because certain features are highly unlikely to occur in some of the contexts. For instance, the discourse particle *like* was only included in 'spoken informal' and 'spoken formal', because it is very unlikely to be encountered in written contexts.[7]

Each sentence had to be rated by respondents on four criteria: 'correctness', 'acceptability', 'goodness', and 'educatedness', using five point semantic-differential scales: CORRECT-INCORRECT, ACCEPTABLE-UNACCEPTABLE, GOOD ENGLISH-BAD ENGLISH, and EDUCATED-UNEDUCATED, as exemplified in Figure 4.2. For spoken stimuli, the survey contained a link to an audio file, followed by the same kind of structured response as the one exemplified in Figure 4.2.[8] The ratings for these four semantic-differential scales were taken as evidence for the ways in which attitudes of speakers towards the use of the six linguistic features differed across a number of variables, explained below. These variables were established on the basis of a number

[7] An exception here might be online language use, or specifically the language used on social media and in discussion groups. However, even in these contexts there is little evidence as to the extent to which the discourse particle *like* would be used. If it does occur, this is likely a fairly new development.

[8] The entire survey is available at https://bit.ly/2xWraST.

Feature	Context	Stimulus sentence
ain't	Spoken informal (conversation)	I ain't going to see them next month.
ain't	Spoken formal (interview)	In school they ain't pushing me, they are encouraging me.
ain't	Written formal (newspaper article)	You won't move forward in your career if you ain't brave enough.
like	Spoken informal (conversation, male)	Didn't you, like, all like go to, erm..., like a boot camp?
like	Spoken informal (conversation, female)	I've like done a couple of like summer camps in like languages and accounting.
literally	Spoken informal (conversation, male)	I literally died from boredom on my date last night!
literally	Spoken informal (conversation, female)	There is story in this book that literally blew my mind!
literally	Written informal (social media)	This book literally blew my mind.
negative concord	Spoken informal (conversation)	I'm strong minded and I'm not going to let nobody lead me off in the wrong direction.
negative concord	Written informal (text message)	I'm sorry. But I'm not going to argue with nobody.
negative concord	Written formal (novel)	I thanked the good lord that I had not killed nobody.
pronouns: object *I*	Spoken informal (conversation)	I think this has been the trouble between you and I.
pronouns: object *me*	Written formal (academic article)	These findings have been very important for my colleagues and me.
pronouns: object *I*	Written informal (social media)	This trip has been a great adventure for my parents and I.
pronouns: object *I*	Written formal (professional email)	The collaboration with your company has been a great pleasure for my workers and I.
pronouns: subject *me*	Spoken informal (conversation)	Me and my husband went to a party with several other young couples.
pronouns: subject *I*	Written formal (professional email)	My colleagues and I will look into this and get back to you as soon as possible.
pronouns: subject *me*	Written informal (text message)	Me and dad are on our way home!
pronouns: subject *me*	Written formal (professional email)	My team and me are working to resolve your problem as soon as possible.
split infinitive	Spoken formal (radio interview)	So, I would encourage young men and women to seriously consider a career in law enforcement.
split infinitive	Written informal (social media)	Trying to decide if there is anything interesting to further explore in my new town.
split infinitive	Written formal (magazine article)	This therapy has been shown to significantly reduce the risks of heart attacks and strokes
split infinitive	Written informal (social media)	Trying to find out if there is anything interesting to explore further in my new town.

Table 4.11: Stimuli sentences used in the survey

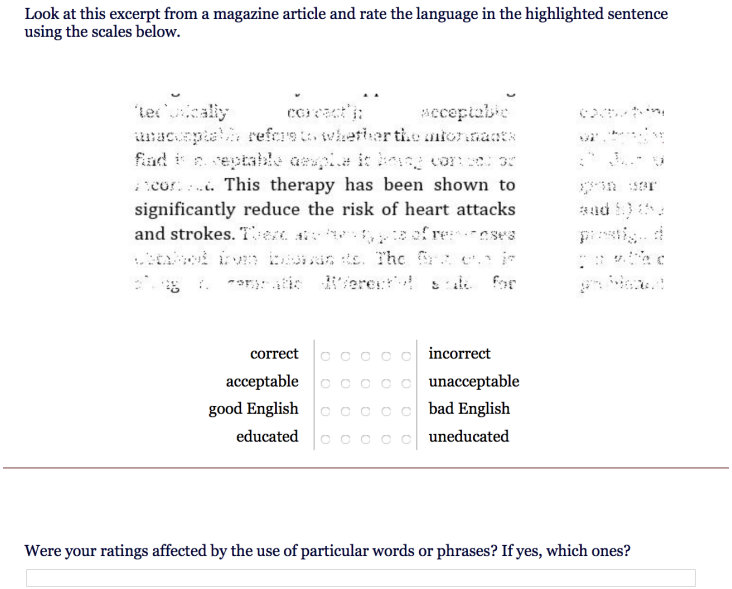

Figure 4.2: An example of the way stimuli sentences were presented to participants in the survey

of general questions about the respondents' backgrounds, which were also included in the survey. The second part was a follow-up unstructured interview with each of the respondents, after the completion of the sentence evaluation. The main purpose of the interview was to allow respondents to reflect on the survey, as well as to communicate thoughts and observations they may have felt were impossible to address in the survey. The interviews were thus fairly unstructured, but the topics covered were naturally related to the respondents' attitudes to language use, as well as the usage problems covered in the survey.

The variables included in the analysis of attitudes are the following. The dependent variables are the ratings of the stimuli sentences in different contexts. The three independent variables are: AGE, GENDER, and ETHNICITY, established on the basis of relevant questions in the survey. Age was operationalised as a nominal variable with two levels 29 AND BELOW and 30 AND ABOVE. The information about the gender of the respondents was obtained by asking an open question ("What is your gender?"). This produced binary data: all of the respondents chose either MALE or FEMALE. Consequently the gender variable was operationalised as a binary variable, although there is an increasing tendency in sociolinguistic research to operationalise

it as a categorical variable, i.e. with multiple levels, such as *male*, *female*, and *other*. Information about ethnicity was similarly obtained with an open question ("How do you describe your ethnic background?"). On the basis of the answers, the respondents were grouped into three groups in terms of ethnicity; this resulted in a categorical variable with three levels: AFRICAN AMERICAN, OTHER, and WHITE. The effects of these variables on the ratings of the stimuli sentences will be explored using non-parametric tests for inter-group comparisons of the ratings of the stimuli sentences. The ratings produced ordinal data which are not normally distributed, so testing between the ratings of two groups was done with the Wilcoxon (Mann-Whitney) test for independent samples (Levshina 2015: 108–113). Because multiple comparisons were conducted on the same data, the significance level was Bonferroni corrected (Levshina 2015: 181), and differed for each of the features, as a different number of tests were conducted for each feature. These aspects of the analysis are addressed in more detail in Chapter 7, which discusses the results of the analysis of the speakers' attitude data.

4.6 Conclusion

In this chapter I have presented the general approach to my analysis of the relationship between prescriptive attitudes to usage in American English, patterns of actual language use, and speakers' attitudes. For each of these perspectives, I have outlined the data used, how the data were collected, and the analytical approaches used to explore these data. In the following chapters, the analysis of each of these datasets is presented and discussed.

With respect to the analysis of the data, I have used a number of software programs and tools. I mention each specific tool in the chapter where I discuss the analysis for which I have made use of that tool. To provide a brief overview of these tools: for the analysis of the usage guide data I used the Hyper Usage Guide of English database (Straaijer 2015) and the 'brat' annotation tool (Stenetorp et al. 2012). I used R (R Core Team 2013) for all statistical analyses and visualisations. `Ggplot2` (Wickham 2009) was used to produce most of the plots in Chapters 5 and 6 and the `Likert` package (Bryer and Speerschneider 2017) was used to produce the plots in Chapter 7. The package `rms` (Harrell 2018) was used to conduct logistic regression analyses.