16

COPY EDITORS, (NOT) ALL ALIKE

Morana Lukač and Adrian Stenton

1. Introduction

1.1 Copy editors and proofreaders

Until recently, the work of copy editors and proofreaders had hardly been a subject of attention for linguists conducting empirical research. When mentioned at all, it has been considered primarily anecdotally and described as an effort to ensure uniformity in language use by suppressing variation and, in effect, enact the "ideology of language standardization" (Milroy & Milroy, 2012, p. 68). One rare and often-invoked account of copy-editing is laid out in Deborah Cameron's theory of verbal hygiene first published in 1995 in the book of the same name (see also Cameron, this volume). Verbal hygiene broadly encompasses people's attempts to "clean up" language so that it would "conform more closely to their ideals of beauty, truth, efficiency, logic, correctness and civility" (Cameron, 2012, p. vii). Among an array of different verbal hygiene discourses and practices described in the book is copy-editing, which, in Cameron's account, becomes synonymous with enforcing the rules included in publishers' style guides. Whilst dealing with texts that are mainly already written in standard English, Cameron further argues, copy editors engage in the process of hyperstandardizing texts and removing variation from the few marginal grammatical contexts where it exists (2012, pp. 47, 53). Consequently, theirs is a crucial role in maintaining the illusion of the standard language. We will argue that, although without question informative, Cameron's account is far from complete. In academic text production, which is central to our study, we will show that copy editors do far more than engage solely in hyperstandardization, but rather mediate text production in a number of ways.

Our research endeavour resonates with the reorientation towards studying prescriptivism as a relevant sociolinguistic factor, and we aim to contribute to the work of scholars who, more recently, have directed their gaze towards the "coal-face of standardization" (McArthur, 2001, p. 4). Jonathon Owen (2020) and Linda Pillière (2020) have both independently made efforts to capture the patterns and variation found in copy-editing practices. Owen (2020) did so by comparing the types of changes novice and experienced copy editors introduced to selected texts. To return to the point made above, Owen's findings challenge Cameron's claim that ensuring prescriptively correct usage is a primary concern for copy editors. Their practices are more complex than had previously been assumed, he argues. Consistency clearly outweighs the

DOI: 10.4324/9781003095125-18

prescriptive notions of correctness as a criterion for introducing changes to texts. Pillière (2020) aimed to uncover differences between British and American copy editors by zeroing in on four linguistic features which go beyond dialectal differences in spelling, punctuation, and lexis. By looking at pronoun use following comparative *than*, the use of *one another* and *each other*, the passive voice, and existential *there*, Pillière sought to find out whether the differences between American and British English copy-editing practices inform us about their respective values and the state of prescriptivism today. Although she did not observe clear trends separating one group of editors from the other, Pillière, like Owen, did find that "copy editors do not form a monolithic group" and that their decisions "vary not just along national lines, but within the various age categories" (2020, p. 288).

In an attempt to answer the call for further research formulated by both of the cited authors, we surveyed 288 copy editors and proofreaders based all around the English-speaking world (see below), and asked them to edit six short excerpts, all of which were taken from the Stenton Corpus, a 12-million-word corpus of international academic English texts. All of the excerpts included the noun *data*, which, although traditionally seen as a plural noun, has been increasingly used as a singular construction. Our first aim was to establish whether copy editors recognize singular and plural *data* as two distinct usages depending on the context in which the noun occurs. We also set out to explore whether there is variation across speakers of different varieties of English and across different age groups in their treatment of *data*. The two sociolinguistic variables should point to regional differences and apparent-time change towards more pervasive usage of singular *data* agreement in academic texts. Finally, our qualitative analysis of the arguments that copy editors and proofreaders introduce for explaining their decisions provides a starting point for understanding the shared values of this community of practice in their complexity.

1.2 Data is/are

The construction of *data* as a singular has been firmly entrenched in usage debates from the beginning of the twentieth century onwards. The *Guardian* newspaper, which often consults members of the public on writing guidelines (Lukač, 2018, p. 118), has referred to the usage of *data* as a "contentious issue" and one which evokes polarised discussions on Twitter and among style guide authors alike (Rogers, 2012). According to the same article by Rogers, whilst the *Guardian* style guide author David Marsh finds the plural usage to be "hyper-correct, old-fashioned and pompous," style recommendations in institutions such as the Office for National Statistics and Royal Statistical Society maintain that plural usage is preferred. In the Hyper Usage Guide of English (HUGE) (Straaijer, 2014), a large albeit not exhaustive database of 77 usage guides published in English between 1770 and 2010, Latinate plurals are first commented upon in an American publication, Joseph Fitzerald's *Word and Phrase: True and False Use in English* (1901). His entry reads as follows: "Datum, data; from which it is seen that data is plural; it is sometimes ignorantly – that is to say, by those who don't know Latin – taken to be singular." From today's perspective, Fitzgerald's pronouncement gives away its age at first glance. Although Latin legacy plural forms have continued to be one of the clearest staples of "restorative prescriptivism" well into present day (Curzan, 2014, p. 24), recent corpus-based analysis of the usage demonstrates that the singular *data* construction outnumbers the plural one with a ratio of 3:1 across inner circle varieties of English (Peters, 2018, p. 48).

Due to the increased frequency of its usage, especially in registers related to science and computing, the noun *data* has undergone a semantic extension. Both the singular and the plural construction are codified and, according to the *OED*, they appear in semantically distinct contexts. In scientific writing, *data* is treated as a plural noun when (i) referring to items of (chiefly numerical) information, typically collected for reference, analysis, or calculation, whereas it is used as a mass noun when (ii) related items of information are considered collectively:

> (i) We often find that no **data have** been fabricated
>
> *OED, s.v. datum, n.*

> (ii) **Data** on long-term effects on healthy users **was** not yet available.
>
> *OED, s.v. data, n.*

Although the singular *data* construction in (ii) continues to be discussed within the prescriptive canon (cf. Heffer, 2010, pp. 53–54; Taggart, 2010, pp. 40–42), the high level of acceptability of *data* as a singular construction among speakers of British English had been demonstrated as early as half a century ago in an attitude survey conducted by Mittins, et al. (1970, pp. 13, 30–32). By the time Ebner carried out a comparative study to that of Mittins et al. in 2017 (pp. 149–151, 199–213), she considered the plural usage to be more salient than the singular one. Rather than including *data is* in her survey stimulus sentence, as Mittins et al. did (*The **data is** sufficient for our purpose.*), Ebner opted for *data are* (*The **data are** often inaccurate.*). Strikingly, the acceptability ratings for *data are* in 2017 were, with an average acceptability of 48.5 per cent, 20 points lower than those for *data is* in 1970 (69 per cent).

Both Ebner (2017) and Peters (2018) come to a similar conclusion when they observe that the proscriptions against the singular usage of *data* demonstrate "the distinction between norms and customary usage" (Ebner, 2017, p. 213) as well as "the strength of linguistic tradition, and the slow adaptation of legacy Latin forms in modern English" (Peters, 2018, p. 41). As pervasive as *data* as a singular collective may be in usage, in examining the GloWbE (Davies, 2013) and the COCA (Davies, 2008) corpora, Peters clearly demonstrates that the last stronghold of the word's plural usage is in the academic registers of North American English (2018, p. 46). Moreover, authoritative publications such as the *7th Edition of the Publication Manual of the American Psychological Association* still prescribe the plural usage of *data* (APA, 2020, p. 162), as do authors of more conservative usage guides.[1]

Given the high acceptability ratings for singular *data*, even, to return to Ebner's example above, in contexts where plural *data* would be expected according to the *OED*, the question arises whether and to what extent the prescription against its usage is still enforced. In this chapter we aim to test the assertion that singular *data* is an example of a stark contrast between usage, which favours it, and the prescriptive view that it should be replaced with a plural construction, particularly in academic registers. To examine this usage–prescription dichotomy more closely, we set out to explore the work done by the copy editors and proofreaders whose interventions are part and parcel of the publishing process. As institutional gatekeepers, their decisions directly shape (written) standard English, and are telling of what is deemed acceptable in published texts. Analysing the edits of *data* agreement can help us decide whether we are observing a change in progress. In other words, our investigation aims to shed light on whether the plural construction continues to be preferred in academic registers.

2. The survey

Our study involved a survey of copy editors and proofreaders for the purpose of determining the differences in the treatment of *data* agreement depending on the context in which the noun occurs, as well as the respondents' age and variety of English. Over the period May to June 2020, we sent emails to 13 organizations for copy editors and proofreaders based in Australia, Canada, Hong Kong, India, Ireland, the Netherlands, South Africa, the UK, and the US, together with three international organizations. All but one forwarded the survey to their members. Through our snowball sampling method, we further recruited respondents through Twitter, blogs, and LinkedIn. While 622 people began answering the survey set up through the online survey software tool Qualtrics, only 288 respondents completed it fully, with most of those dropping out citing time constraints. The first part of the survey comprised a section with questions eliciting sociodemographic data, including the respondents' age, variety of English, education level and field, length of editing experience, the type of texts usually edited, and sources consulted on questions of English style and grammar. The distribution of the respondents across age groups and variety can be seen in Table 16.1. In terms of their education level, the participants formed a largely homogeneous group with 91% stating that they were university educated. The three most common fields of editing included arts and humanities, business and management, and social and behavioural sciences. Our analysis showed no significant differences in responses depending on education, field, and length of editing, and the references consulted. Some interesting although not statistically significant patterns did emerge, however, when we considered the references mentioned by our respondents, and we return to these below.

While the youngest among the respondents was 19 and the oldest 78, the mean age was 49. Nearly half of the respondents were speakers of US-American English (46.5%), followed in number by British English speakers (18.4%), non-native (13.2%), South African (9%), and speakers of Canadian (7.3%) and Australian English (3.5%). The smallest groups included speakers of Irish, Northern Irish and Philippine English varieties. One 77-year-old speaker defined their variety as 'UN/international English, UK spelling', coded as International in Table 16.1.

In the second, editing part of the survey, we presented six short texts all of which included the use of the word *data* in different contexts. Each raised different issues with the use of *data*, and these will be detailed in the Rationale section. Respondents were asked to "Click and highlight the parts of the text (if there are any) in the […] example that, in your opinion, require editing". They were then asked to provide their proposed edit, and to add any comments. The six texts came from the Stenton Corpus, which will be described below. The texts also raise issues of presentation, in that the examples were set in a short context, rather than as single sentences or phrases, with the target uses highlighted, which has been seen as a contentious issue in other surveys. These two issues will be discussed in their respective sections below.

Table 16.1 Sociodemographics of the survey respondents

Age	*19–29*	*30–39*	*40–49*	*50–59*	*60–69*	*70–78*	*Total*
	20	64	71	58	52	23	288
Variety		Native			Non-native		Total
		250			38		288
	American	Australian	British				
	134	10	53				
Region	Canadian	International	Irish				
	21	1	2				
	N. Irish	Philippine	S. African				
	1	2	26				

3. The Stenton Corpus

The corpus being used for this study, the Stenton Corpus, is what McEnery and Hardie term an "*opportunistic* corp[us]", in that it consists of "nothing more nor less than the data that it was possible to gather for a specific task" (2012, p. 11). The corpus consists of 1031 manuscripts (mss) accepted for publication in three Law journals and three Language journals,[2] published by Cambridge University Press (CUP), in Cambridge, England, over the period 2006 to 2016.[3] The total word count of the corpus is 11.58 million: the Law journals contain 2.58 million words, and the Language journals contain nine million words. The most pertinent aspects of the Stenton Corpus for this study are that it consists of manuscripts that have not been copy-edited and that it does not reflect a single regional variety of English. The Stenton Corpus forms the basis of a separate study (Stenton, in progress), but was also the source of the extracts used in this survey.

3.1 The manuscripts

The 1031 manuscripts in the Stenton Corpus are not edited, in the sense that they are not copy-edited. The mss have all been reviewed by the journals' editorial boards, they have been sent out for blind peer review, and they have been revised. Once the mss are approved, they are sent out for copy-editing, and for subsequent proof-reading and proof collation. The versions of the mss used in the Stenton Corpus are thus the unedited mss as received from CUP, and have not, to our knowledge, been professionally copy-edited. This lack of copy-editing is thus potentially a major difference between the Stenton Corpus and many other corpora of written English. The significance of this aspect of the Stenton Corpus is that it avoids what Rawlins and Chapman (2020, p. 10) refer to as "one of the weaknesses of corpus research – many of the texts in the corpora have been edited, thereby giving the attitudes and practices of copy editors an outsized influence in the published language". It is also why they are such a useful resource for the current study.

3.2 The authors

There were 1657 different authors listed for the 1301 mss. The only information that is available about the authors is their institutional affiliation by country at the time the ms was submitted. There is no information about the nationality, age, or gender of the authors, and none on native languages. For the Stenton Corpus as a whole, four of the seven native English-speaking areas (Trudgill and Hannah, 2017, p. 12) provide the highest number of authors: the US (686), the UK (309), Canada (149), and Australia (143). New Zealand provides 11 authors, Ireland 8, and South Africa 6. The eight "second-language varieties of English" (2017, pp. 128–145) ESL countries provide between them a total of 52 authors, with Singapore providing the bulk of those at 36. The seven native English-speaking areas thus provide 59 per cent of the authors by affiliation.

Notwithstanding the fact that all of the mss were published in England, given the lack of detailed information about the authors, and the wide range of country and institutional affiliations, we cannot assign the mss to the variety of British English. Instead, in the spirit of Trudgill and Hannah (2017), we have chosen to label the language of the Stenton Corpus 'International Academic English'.[4]

3.3 Context

We decided to present the examples in context, typically including one sentence before and after the one containing the use of *data*. Context is generally not included in survey questions, despite it being discussed in both usage guides (see e.g. Gilman, 1989, pp. 23, 122; Peters, 2004, p. 138; Sayce, 2006, p. 25; Taggart, 2010, p. 43) and usage studies (see e.g. Ebner, 2014, pp. 3–4; Tieken-Boon van Ostade & Ebner, 2017, §4.4). Usage studies have also found that respondents call for more context in making their decisions on acceptability (Pillière, 2018, p. 262; Tieken-Boon van Ostade, 2020, p. 167), and we wanted to discover if usage beyond the immediate sentence had any influence on attitudes to the usage in question. Our assumption is that, with the context of the example being shown, the respondents might be more likely to find an example acceptable.

3.4 Highlighting

We highlighted the usage in question within the example, following Mittins et al. (1970). This use of highlighting to identify the usage is also contentious, however. Tieken-Boon van Ostade (2013) has pointed out that one of the difficulties with highlighting the phrase of interest is that the respondents "would be biased against features which they knew, however dimly, to clash with accepted standard practice" (2013, p. 4). This notion of bias is taken up by Tieken-Boon van Ostade and Ebner (2017), and by Ebner (2017, 2018), with Ebner favouring "the methodological advantage of not highlighting the usage problems" (2018, p. 148).

Tieken-Boon van Ostade (2013) noted that many of her respondents "failed to see [the usage] as a potential usage problem" (2013, p. 7), given that the form was not highlighted. In reporting on the same survey in 2020, Tieken-Boon van Ostade again comments that "many people … failed to identify what usage problem they were asked to comment on. There would therefore be something to be said for highlighting the issue tested after all" (2020, p. 168). Kostadinova (2018) has pointed to the danger of discovering only "the attitudes speakers think they are expected to have" (2018, p. 208) if the usage is highlighted. Meanwhile, Ebner (2017) added that "consciously highlighting the investigated items no longer seems to fit the contemporary research undertaking as awareness [i.e., of the problem being investigated] is becoming an increasingly important factor" (2017, p. 111).

For this study, one of the factors contributing to the decision to highlight the phrase of interest was that the Stenton Corpus, the source of the examples, is made up of texts which have not yet been copy-edited, and so were more likely to contain what could be seen as errors in addition to the phrase of interest, especially given that they were presented in a longer context. In part because the topic of interest was so specific, and in part because it was presented in (an un-copy-edited) context, we concluded that it was more helpful to highlight the phrase than not.

4. The six extracts

Our respondents were presented with the following extracts:

> *Extract 1 [IJC]*
> Indeed, estimates suggest that only about 10% of Nigerians have access to essential drugs – a figure that presumably includes the over 2.7 million people living with

> HIVAIDS. With regards the rate of doctors per citizens, recent **data** also **suggests** [suggest] that less than 30 doctors are available to serve about 100,000 people. Indeed, the failings of the public health systems have seen Nigerians resort to private health services, which are estimated to provide 65.7% of the nation's healthcare needs.

This is the simplest of our extracts, in which we expected those respondents who favour the plural use of *data* to change *suggests* to *suggest*.[5]

> *Extract 2 [LCO]*
> Some of **this** [these] **data is** [are] already in analysis-friendly form, such as social network information (Lewis, Kaufman, Gonzalez, Wimmer, & Christakis, 2008; Lerman & Ghosh, 2010), diurnal activity patterns (Krishnamurthy, Gill, & Arlitt, 2008), reputation (Standifird, 2001), or Facebook 'likes' (Kosinski, Stillwell, & Graepel, 2013). An enormous **amount** [number] of **data**, however, **is** [are] in the form of human generated text, and that is not something that can be directly analyzed. Despite the difficulties of using computer algorithms for analyzing written text, the field is quickly developing.

Extract 2 is more complex than Extract 1, although it mimics it on the surface, in that, again for those respondents who favour the plural use of *data*, we would expect the following edits:

(i) *data is* to *data are*;
(ii) *data, however, is* to *data, however, are*.

For (i), we would also expect *this* to *these*:

(iii) *Some of these data are*.

For (ii), there are alternative analyses of the subject noun phrase *An enormous amount of data*, such that (iv) *data* functions as the head, with *An enormous amount of* being a pre-modifier, and that (v) *amount* functions as head, with *of data* being a post-modifying prepositional phrase. In the case of (iv), *data* would determine the number of the verb, so the result would be:

(iv) *An enormous amount of data, however, are*.

In the case of (v), *amount* would determine the number of the verb, so the result would be:

(v) *An enormous amount of data, however, is*.

Further, for those who prefer (iv), *amount* could be changed to *number*, to reflect the change from singular/mass to plural:

(vi) *An enormous number of data, however, are*.

Option (vi) would not, of course, be open to those who regard *amount* as the head of the subject noun phrase.

> *Extract 3 [JCL]*
> To assess the effect of Age on the production of scrambling, the **data were** [was] adjusted using the following procedure. The two-year-old group was not included in the ANOVAs analysis because there were too few participants (N=6) in it and a high percentage of **data** (14.6%) **was** missing. The analysis of the remaining 4 age groups was based only on the **data** from 10 participants (the number of three-year-old children) per group.

Extract 3 is similar to Extract 2, in that the first use of *data* would not be changed by those who prefer the plural, so it is the first opportunity for those who prefer the singular to make a change:

(i) *the data was adjusted*.

For those who prefer the singular, the second use of *data* would also not be changed, irrespective of how they saw the head of the subject noun phrase, i.e., as *data* or as *percentage*. However, for those who prefer the plural use of *data*, there would be two options, again depending upon how they view the head:

(ii) *a high percentage of data were missing*;
(iii) *a high percentage of data was missing*.

It is because of options (ii) and (iii) that we encouraged respondents to explain their decisions, as (iii) could be acceptable to both groups, depending on their analysis of the subject noun phrase. We did not expect any changes to the third use of *data*.

> *Extract 4 [JCL]*
> While the above discussion does not exhaust the range of theories entertained in the literature, it is enough to demonstrate that current empirical **data is** [are] consistent with a wide range of possibilities, and also to point out what **kind** [kinds] of **data is** [are] needed to constrain the theoretical possibilities. In particular, now that it is well established that young children have expectations about the semantics of a verb given its syntax, we need to determine what the boundary conditions and constrains on those expectations are. The present work provides some of these boundary conditions and constraints, but **data** from younger children and from additional types of verbs (e.g., contact verbs) **is** [are] needed.

Extract 4 raises the same issues as before, and here we expect to see a strong division between those who prefer plural and those who prefer singular uses of *data*.

> *Extract 5 [JCL]*
> It is essential in a setting with great linguistic diversity (over 40 languages are used in Kenya) that assessment instruments are easily adaptable. Obtaining comprehensive item sets is difficult in a situation where **little** [few] previous **data** on child language use **is** [are] available. In creating the Kilifi CDIs we therefore necessarily started with an English version because there was no closer language version available. Existing **data** available on the languages studied here, Kigiriama and Kiswahili, **suggested** that children are more advanced on some aspects of grammatical development.

In Extract 5, we expect the use of *data* to follow singular or plural preferences, and there is the additional question, for those who prefer the plural use, of whether they will also change *little* to *few*:

(i) *little previous data is available*;
(ii) *few previous data are available*.

We do not expect the second use of *data*, *Existing data … suggested*, to be changed, as the past tense verb is not marked for number.

> *Extract 6 [AJL]*
> Further commentary was provided on the meaning of 'available', which McCaffrey suggested to mean that, 'the notifying State is generally not required to do additional research at the request of the potentially affected State, but must provide only such relevant **data** and information as **have been developed** in relation to the proposed use and **are** readily accessible'. McCaffrey suggests that where **data** or information **is** [are] not readily available, but **is** [are] accessible only to the notifying State, 'it would generally be appropriate for the former to offer to indemnify the latter for expenses incurred in producing the requested material'. The Watercourses Convention accordingly obliges the notifying State to cooperate with the notified States to provide them, on request, 'with any additional **data** and information that **is** [are] available and necessary for an accurate evaluation'.

Extract 6 raises a number of further issues. First of all, the first and third uses of *data* are within quotations, and so we would not expect them to be revised at all. The second use, *where data or information is […]*, presents a new problem for the copy-editors, that of the compound subject, *data or information*. Here, the proximity principle might suggest that the verb remain singular, as *information* would be very unlikely to appear with a plural verb. There is also the almost mirroring of the quotation that includes the third use of data, *any additional data and information that is […]*. Here, though, the compound subject is conjunctive, *and*, rather than disjunctive, *or*, but the quoted author nonetheless uses a singular verb, notwithstanding their use of the plural *have been developed* with a similar conjunctive subject in the first quotation. This we found to be an interesting problem for our copy editors in terms of consistency. It should be noted that, in the respondents' comments on these extracts, we were not surveying the editors' grammatical knowledge, we were simply asking them to explain their copy-editing decisions in whatever way they chose.

5. Quantitative analysis

5.1 An overview of highlighting and editing choices

The survey software tool Qualtrics enables the incorporation of highlighted questions, which allows researchers to present survey participants with an interactive text sample. Participants could select words from the text presented to them and evaluate them. Our criterion for highlighting parts of the text was whether or not it required editing. The distribution of highlighted examples of *data* agreement per example are illustrated in Figure 16.1.

In all but one example (E3 *data were adjusted*), to which we will return below, in varying degrees, most editors chose not to highlight the examples.

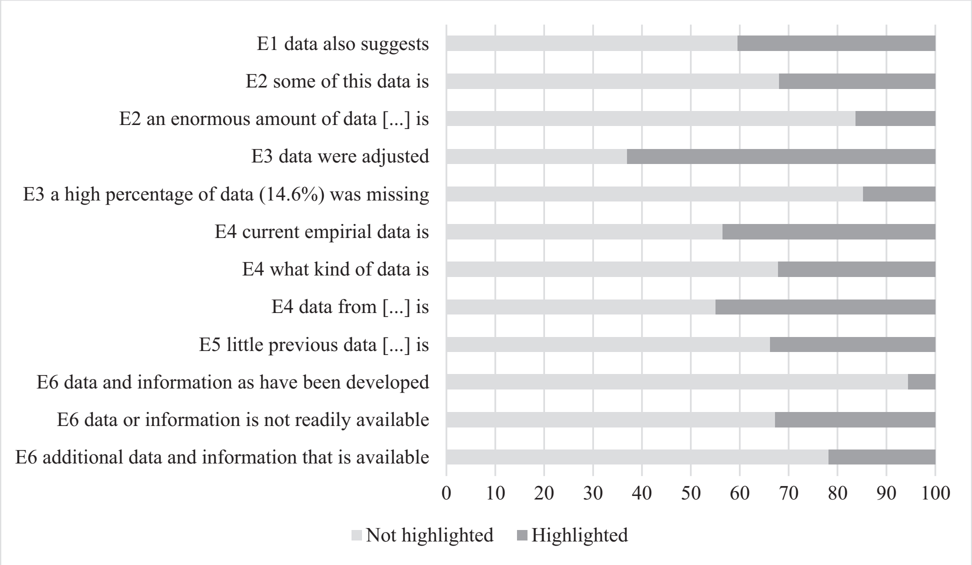

Figure 16.1 Percentage of highlighted data agreement across six extracts.

After instructing them to highlight parts of the text that they deemed problematical, we asked our respondents to propose an edit of the highlighted texts. All of the edits, or lack of any intervention, fell into one of five categories: the respondents chose singular agreement by either changing the verb or not intervening with the original text (Singular), they opted for singular agreement and revised the adjacent text (Singular revised), they chose plural agreement by either changing the verb or not intervening with the original text (Plural), they opted for plural agreement and revised the adjacent text (Plural revised), or they revised the original text and avoided *data* agreement issues altogether by changing the tense of the verb or replacing the noun (Other revised). The percentages of the proposed *data* agreement edits per category are illustrated in Figure 16.2.

Although highlighting and the proposed edits largely overlap, there is one obvious discrepancy in the E3 *data were adjusted* example. Whereas 63 per cent of the respondents highlighted data agreement in the respective example, only 33 per cent edited it when asked to do so. We propose that acquiescence bias may be at play here, with the respondents generally avoiding editing the text presented to them. This may be because they engaged in light editing in this survey and avoided introducing changes where possible, for time efficiency or lack of investment. Thus, in identifying patterns in the responses, we chose to focus, in the remaining parts of this chapter, only on those 112 respondents who had answered the survey in full.

5.2 Editing patterns

The two-step cluster analysis procedure is an exploratory tool designed to reveal natural subgroups (or clusters) within a larger sample that would otherwise not be apparent. In our study, choosing this approach seemed to be the most promising in determining whether our respondents provided consistent responses and formed groups in terms of how they chose to edit *data* agreement across the six extracts presented in the survey. The question was then

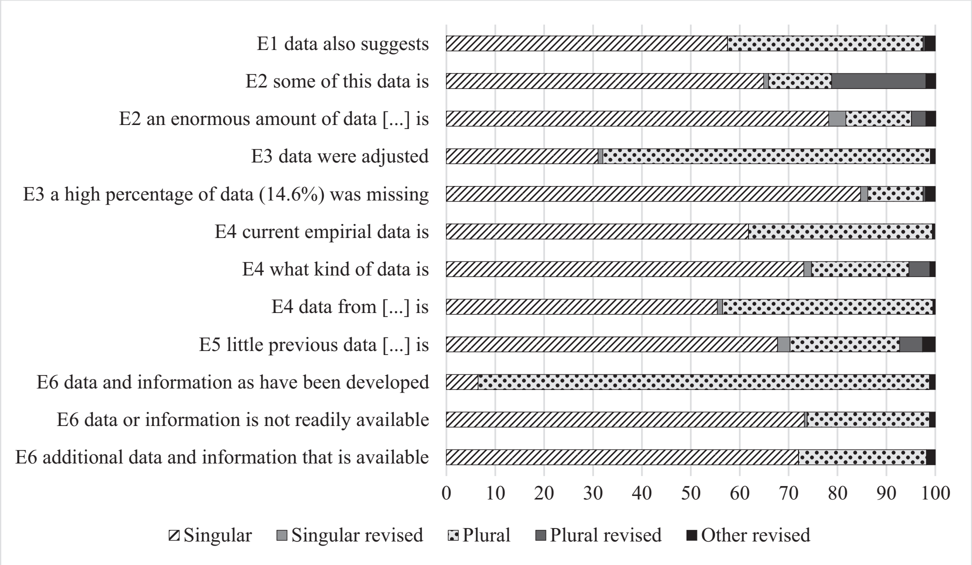

Figure 16.2 Percentage of proposed edits of data agreement.

whether we can speak of "profiles of editors" in terms of whether they opted for singular or plural verb agreement with the noun *data*. We thus merged the above-mentioned categories Singular revised and Singular and Plural revised and Plural into two categorical variables (Singular and Plural) of the edits of verbs in all extracts and performed a two-step cluster analysis implemented in the base package of IBM SPSS Statistics (version 27.0). We allowed for the procedure to automatically determine the best number of clusters emerging from our dataset. It is worth noting that cluster analysis generally does not allow for missing values, which meant that we could only perform this analysis using the data from the 112 respondents who answered all of the questions posed in the survey and actively edited all of the paragraphs. The cluster quality proved to be fair (average silhouette or ASW equalled 0.5 on a scale from −1 to +1, indicating that the object is well matched to its own cluster and poorly matched to neighbouring clusters).

The analysis showed the presence of two clusters. Of the 112 respondents, approximately half favoured plural agreement (Group 1: 55 or 49.1%) and the other half generally favoured singular agreement (Group 2: 57 or 50.9%). Not all the examples of *data* agreement from the survey were equally good predictors of whether a respondent would belong in Group 1 or Group 2. And, more importantly, although Group 1 tends to generally favour plural and Group 2 singular, this did not mean that they consistently chose singular or plural across the six extracts. For example, although Group 1 predominantly chose plural agreement, in E4 *what kind of data is*, the most common response (43.9%) was not to edit, but to keep singular agreement.

Some of the examples of *data* agreement from the survey indicate starker divisions than others. Table 16.2 lists the edits per paragraph, from the one which best indicates whether someone opts for plural or singular agreement to the one that does so the least. In the parentheses next to each example, the percentage of the preferred response in the group is indicated. This means, for instance, that 96.5 per cent of the respondents in Group 1 chose to edit the third example from E4 *data from […] is needed* and change it to plural agreement, whereas 96.4

Table 16.2 Clusters per predictor importance

Group 1	*Group 2*
E4 *data from […] is needed* (plural = 96.5%)	E4 *data from […] is needed* (singular = 96.4%)
E4 *current empirical data is* (plural = 91.2%)	E4 *current empirical data is* (singular = 100%)
E4 *some of this data is* (revise plural = 52.6%)	E4 *some of this data is* (singular = 98.2%)
E5 *little previous data […] is* (plural = 50.9)	E5 *little previous data […] is* (singular = 98.2%)
E1 *data also suggests* (plural = 86.0%)	E1 *data also suggests* (singular = 74.5%)
E3 data were adjusted (plural = 98.2%)	E3 *data were adjusted* (singular = 50.9%)
E4 *what kind of data is* (singular = 43.9%)	E4 *what kind of data is* (singular = 49.5%)
E2 *an enormous amount of data […] is* (singular = 57.9%)	E2 an enormous amount of data [...] is (singular = 98.2%)
E6 *data or information is not readily available* (singular = 54.4%)	E6 *data or information is not readily available* (singular = 83.6%)
E6 *data and information as have been developed* (plural = 96.5%)	E6 *data and information as have been developed* (plural = 90.9%)
E3 *a high percentage of data (14.6%) was missing* (singular = 0.2%)	E3 *a high percentage of data (14.6%) was missing* (singular = 92.7%)
E6 *additional data and information that is available* (singular = 59.6%)	E6 *additional data and information that is available* (singular = 78.2%)

Table 16.3 Group 1 and Group 2 per age category

Age	*Group 1*	*Group 2*	*Total*
19–29	1	5	6
30–39	14	11	25
40–49	10	14	24
50–59	14	9	23
60–69	8	15	23
70–78	10	1	11
Total	57	55	112

per cent of Group 2 decided to keep singular agreement. The percentages in parentheses refer to the most common category of responses for the cited example per group.

5.3 Distribution per age and variety

We were interested in whether group membership (Group 1 or 2) is associated with the respondents' age or variety of English and for that we used the Chi-square test implemented in SPSS. The results of the test revealed no significant relationship between group membership and age ($\chi^2(5, N = 112) = 16.2, p = .014$), nor between group membership and variety ($\chi^2(6, N = 112) = 9.9, p = .128$). However, some interesting patterns did emerge. If we take a closer look at the youngest and the oldest group in the sample in Table 16.3, it becomes clear that 19–29-year-olds favour singular agreement, with the opposite trend among those aged 70–78. Although our sample is opportunistic and small, this finding tentatively confirms what has been argued in diachronic analyses elsewhere, namely that we are observing apparent-time change in favour of singular over plural agreement.

Table 16.4 Group 1 and Group 2 per variety

Variety	*Group 1*	*Group 2*	*Total*
American	19	26	45
Australian	4	2	6
British	11	17	28
Canadian	3	3	6
Non-native	8	4	12
Philippine	1	0	1
South African	11	3	14
Total	57	55	112

The summary in Table 16.4 shows that most US-American and British copy editors chose singular agreement, albeit with only a slight majority. The opposite is true of Australian, non-native, and South African respondents. There may be several different explanations for this, and here we offer one by evoking the "cultural cringe" concept. Severin and Burridge (2020) introduce it in their discussion of prescriptivism in Australian English, whose speakers, the authors claim, "live with a proverbial chip on their shoulder, constantly comparing themselves to other (often Anglophone) countries, primarily Britain and the USA" (2020, p. 204). Such linguistic insecurities, we argue, are not relevant in the context of Antipodean varieties, but can generally be applied to non-native English-speaking areas. This group of English speakers may see their own everyday usage as further removed from registers such as academic writing and are thus less likely to refer to general usage as a criterion guiding their decisions, but rather resort to the values shared within the prescriptive canon. The speakers from the two most powerful native-speaking areas might not envisage this gap to be as wide.

6. Qualitative analysis

Having presented the results of our quantitative analysis, in this section we turn to the qualitative investigation of the editing choices. As mentioned above, an interesting feature of the two groups identified through the two-step cluster analysis is that they did not always favour singular or plural in each extract. In this section we attempt to investigate why this should be. For an explanation of what we expected to find, please see the Rationale section above.

Extract 1

The relevant sentence in this extract was:

(i) *With regards the rate of doctors per citizens, recent* ***data*** *also* ***suggests*** *[suggest] that less than 30 doctors are available to serve about 100,000 people.*

The expectation here, especially with Group 1 respondents (i.e. those who favoured the plural), is that they would edit *recent data also suggests* to *recent data also suggest*, and 49 of the 57 respondents did exactly this. Of those 49, 16 also made a comment to the effect that "data is plural".

For the Group 2 respondents (i.e., those who favoured the singular), the expectation is that they would make no change to *recent data also suggests*, and of the 55 respondents, only eight in fact edited it to *recent data also suggest*. Of those eight, two made no comment, whilst the other six referred to the context of either an academic journal or a style guide, with the occasional personal preference ("if you prefer data as a plural"). Some comments reflected contradictions between the respondents' attitudes to *data* agreement and their actual practices, which were influenced by various factors, such as house style guides, as this quote goes to show: "Also, for my job as copy editor for an arts and entertainment news weekly, I would need to treat data as singular – it grates on my nerves every time".

Extract 2

The foci of our analysis of Extract 2 were:

(i) *Some of* ***this*** *[these]* ***data is*** *[are] already in analysis-friendly form, such as social network information*; and

(ii) *An enormous* ***amount*** *[number] of* ***data****, however,* ***is*** *[are] in the form of human generated text.*

We expected this extract to be much more interesting, and so it proved. Within Group 1, the respondents fell into two main sub-groups: those who changed both verbs to *are* (19/57), and those who changed the first verb to *are*, but left the second verb as *is* (26/57). A third sub-group of 10 respondents either explicitly re-stated the two verbs as *is*, or indicated no change.

The sub-group of 26, who changed the first verb to *are* but left the second as *is*, also provided the reason for their decision not to change the second verb, which centres around the head of the subject noun phrase, and more specifically whether it should be analysed as *data*, as preferred by the 19 respondents, or as *amount*, with *of data* being seen as a post-modifying prepositional phrase: [*enormous amount* [*of data*]], *however*, *is*. One respondent commented on this at length:

> … Contrary to my edits on the previous text, I'd use the singular verb with 'data'. In the second instance of the word's occurrence in this text, in particular, the subject in question is singular ('amount'), so a singular verb is preferred.

This point was explicitly made by five respondents. One further respondent, clearly aware of this possibility, but who nonetheless changed the verb to *are*, justified the decision by analysing *an amount of* as an adjectival phrase.

For the Group 2 respondents, we would expect no edits to *this data is*, and no edits to *An enormous amount of data … is*. However, if the respondents were to revise the text, we wanted to see if, as well as making the verbs plural, they would also revise the determiner (*this* to *these*) in the first use, and the (partitive) noun (*amount* to *number*), as suggested by many usage guides, in the second use. In fact, none of the respondents revised either use, with some again commenting on the variability of singular/plural usage, some on the priority of consistency, some on the priority of the author, and some justifying the singular use by reference to *some* and to *amount*.

Extract 3

A similar approach seems to have been taken with Extract 3 where we focused on:

(i) *the* ***data were*** *[was] adjusted using the following procedure*; and

(ii) *a high percentage of* ***data*** *(14.6%)* ***was*** *missing.*

Here, we might expect the Group 1 respondents to leave *were* in the plural, and to change *was* to *were*, and 14 of them did just that. Of the remainder, 15 explicitly re-stated the forms as they appeared, and a further 26 made no revisions. One of several respondents who explicitly commented said: "high percentage of data ... was – percentage is the subject and thus takes a singular verb".

So again we seem to have a principled, if often unspecified, approach, whereby *data* is not sometimes treated as plural and sometimes as singular by the Group 1 respondents, but where it is sometimes not seen as the source/target of number agreement with the verb, i.e. it is not seen as the head of the subject noun phrase.

For the Group 2 respondents, we might expect them to revise *the data were adjusted* to *the data was adjusted*, and to not change *a high percentage of data … was missing*. In fact, 27 respondents changed *the data were adjusted* to *the data was adjusted*. Only 2 respondents changed *a high percentage of data … was missing* to *a high percentage of data … were missing*; one of them mentioned consistency with the first use and the other did not comment but left the first use as plural as well. As with the Group 1 respondents, several of those in Group 2 referred to the headedness of the second phrase as a reason for not changing the verb. Five respondents commented on the need for consistency in the extract, and a further five mentioned referring to the journals' style guides.

> *Extract 4*
>
> Extract 4 raises similar issues as before, and here we expected to see a strong division in the two groups between those who prefer plural and those who prefer singular uses of *data* in the following parts the text:
>
> (i) *current empirical* ***data is*** *[are] consistent with a wide range of possibilities*;
> (ii) *also to point out what* ***kind*** *[kinds] of* ***data is*** *[are] needed*;
> (iii) ***data*** *from younger children and from additional types of verbs (e.g., contact verbs) is [are] needed*.

We would expect Group 1 respondents to prefer the plural in all three cases. In fact, there was again a more nuanced set of revisions. There is also a suggestion here that the respondents have now got into their stride, having spent the first three extracts working out both what we were looking for and their own views! For the Group 1 respondents, there was an even split between those who pluralized all three uses and those who pluralized the first and the last, but who left the second as is. Again, this was down to their difference in allocating either *data* or *kind* as the head of the noun phrase, with some explicitly commenting on this: "have taken 'needed' as qualifying 'kind' rather than 'data' and left as singular"; " 'what kind of data' takes a singular verb". Also, some further respondents revised the second use to avoid any apparent conflict, e.g., "to indicate the kind of data needed". Still other respondents revised *what kind of data is* to *what kinds of data are*, with one revising *data* to *datum*, again to resolve the conflict. All of these respondents revised the first and third uses to *are*.

For the Group 2 respondents, the vast majority did not highlight any of the uses of *data* to be changed. Two changed the final use of *is* to *are*, and a further two revised the second use to avoid any conflict. One of the two who changed the second use also gave a long explanation of why they were treating the different uses of *data* differently, and a further 11 commented on how and why they were comfortable with singular *data*, with four of those mentioning the influence of the context.

> *Extract 5*
>
> Here, we expected to see revisions in two places in the text:

(i) ***little*** *[few] previous* ***data*** *on child language use is [are] available*; and
(ii) *Existing* ***data*** *available on the languages studied here, Kigiriama and Kiswahili,* ***suggested*** *that.*

For Group 1, 11 respondents made no revisions at all. Of the remainder, most concentrated on the first use of *data*, revising it to *little previous data on child language use are available*. However, a number of respondents were unhappy with *little data are*, with seven changing *little* to *few*, *scant* or *limited*, and five simply deleting *little*, and then using the plural *are*. Two others recognized the problem caused by *little* and left the verb in the singular, sometimes with a comment such as "'little data' is a set form" and "'little data' is necessarily treated as a mass noun (sg)". Although there was no "need" to revise *data … suggested*, 10 respondents preferred the present tense, with all bar one changing it to plural *suggest*. The odd one who changed it to singular *suggests* also changed *is* to *are* for the first use, so this may be no more than a careless error.

Many respondents in Group 2 were also concerned with revising *suggested* to *suggests*, commenting that this was a "formal document" and that the past tense gave the claim "less weight". As expected for this group, most accepted *data* as singular, with only one respondent changing the second use to plural, but with no explanation.

Extract 6

In three sentences in this extract, we expected to see editorial interventions with respect to *data* agreement:

(i) *… the potentially affected State, but must provide only such relevant* ***data*** *and information as* ***have been developed*** *in relation to the proposed use and* ***are*** *readily accessible;*
(ii) *McCaffrey suggests that where* ***data*** *or information is [are] not readily available, but* ***is*** *[are] accessible only to the notifying State*; and
(iii) *the notified States to provide them, on request, 'with any additional* ***data*** *and information that is [are] available and necessary for an accurate evaluation'.*

Here, for the plural group, starting with the second use of *data*, *data or information is… but is*, 21 out of 55 respondents made no changes. Twenty-four changed this to *data or information are… but are*, and a further 10 changed the first *is* to *are* but not the second. It's not clear whether this was intentional or an oversight. For the last use – *with any additional data and information that is available and necessary* – 39 respondents made no changes, as expected. A further nine made no change but added "[sic]" or "[are]" or "[…]" to identify what they saw as the error. Seven respondents revised *is* to *are*, seemingly heedless of the quotation. Those who made no changes justified their decisions by referring to the quotations, or to consistency; some of those who made no change to the non-quoted use mentioned some version of the proximity rule with disjunctive coordination, or simply stated that it was *information* that determined the verb number. As expected, none of the respondents revised *have* in the first, quoted, use.

As expected, for the singular group this was rather more straightforward, with 39 of the 55 respondents making no changes. Perhaps surprisingly, five respondents changed all three instances of *is* to *are*, including the one in the quotation, whilst a further seven revised the final *is*, in the quotation, to *are*. Again surprisingly, but more in keeping with their group allocation, four respondents changed *have* to *has* in the first quotation. Again, respondents commented on the quotations, consistency and author preference. There was very little in the way of grammatical justification.

Overall, then, it would seem that, whilst we can identify two groups – one favouring plural and the other singular *data* – these groups are most apparent only when there are no confounding contextual influences.

7. Respondent references

As part of the survey, we asked respondents to list their preferred reference sources. We wanted to explore a possible correlation between their editing choices and the advice provided in the references they listed. We had many responses, ranging from "colleagues" to "Google", and many respondents listed several sources. We decided to concentrate again on the 122 respondents in Groups 1 and 2, whose preferred resources were, in descending order of frequency:

Chicago Manual of Style (CMoS)*
Oxford*
Associated Press Stylebook (AP)*
Merriam-Webster
American Psychological Association *Manual of Style* (APA)*
Grammar Girl
Hart's Rules*
Cambridge
Fowler
Grammarly
Strunk & White*
Modern Language Association *Stylebook**
American Medical Association *Manual of Style* (AMA)

The frequency range was considerable, with *CMoS* listed 225 times, and AMA 17 times. Clearly, "Oxford" encompasses a range of titles, including dictionaries and usage guides, as do both Merriam-Webster and Cambridge, but generally the different titles within each group follow a similar approach in their treatment of *data*. *Hart's Rules* and Fowler are also published by Oxford. Grammar Girl and Grammarly are both online resources. The sources with an asterisk were also found in Pillière's (2020) study and are clearly widely used. Pillière's study was in part investigating the influence of style and usage guides on British and American copy-editorial decision-making. She found the relationship complex, and subject to many influences (2020, pp. 271–273).

Of the sources used in the current study, most were equivocal in their use of singular/plural *data*, but when context, i.e. formal or academic, was taken into account, more of them suggested that the plural was a safer choice. The one source that stood out for singular was the AP, which has "revised our guidance to say *data* typically takes singular verbs and pronouns in writing". Notwithstanding this, mirroring Pillière's study, we could find no correlation between reference sources and membership of Groups 1 and 2. Indeed, of those respondents who listed *CMoS*, 13 were in the plural group and 14 in the singular group. In practice, although many respondents provided explanations for their choice of singular/plural, only four respondents in Group 1 mentioned a reference source in their choice: three of them cited APA and the fourth cited *CMoS*, Oxford and the English Academy of Southern Africa. It would seem that the problem of the number of *data* is sufficiently common for the respondents to not need to check it.

8. Conclusion

The copy editors and proofreaders in our study differed in how they approached *data* agreement, and we have clearly demonstrated that the variation in their responses was not random, with half of the respondents who completed the survey favouring singular and the other half plural agreement. Although the traditional sociolinguistic variables of age and variety did not help us completely disentangle the factors involved in the decision-making process, inter-varietal differences and age may play a role in favouring singular or plural agreement. Further research including a representative, rather than an opportunistic sample, should help us clarify the relevance of these factors with more certainty. Due to the number of differences we found, even within more homogenous groups, our survey makes an argument in favour of providing longer examples together with their contexts in studies eliciting attitudes to usage.

Based on the variation we found in our sample, we too, like Owen, are driven to the conclusion that copy editors and proofreaders are “not a homogenous group driven by a single set of values” (2020, p. 302). The practices of these language professionals are guided by the values of their field, to which Amy Einsohn, author of *The Copyeditor’s Handbook*, refers as the four Cs of copy-editing: Clarity, Coherency, Consistency, and Correctness (2006, p. 1). To this we can add a fifth C emerging from our data: language-internal Constraints, such as the perceived semantic distinctions in usage. Finally, copy editors and proofreaders have been described in the past as professionals ensuring uniformity in usage. What our study calls into question is the notion of uniformity altogether. Seeing that our survey examples belonged to the academic register, we expected to see consistent maintenance of the conservative tradition (plural *data*) among our group of respondents. Our findings show, however, that this was hardly the case. The editorial decisions were often divorced from prescriptions, even where we would expect them to be followed. The world of global academic publishing is made up of speakers of a great number of different varieties of English, and the norms, even of a seemingly more rigid variety such as written standard English, are renegotiated rather fiercely. With this in mind, we urge sociolinguists to investigate further how norms are understood and implemented among different groups of gatekeepers. Studies of this sort will help us understand the processes which de facto shape and direct the development of standard written English.

Notes

[1] Simon Heffer’s usage guide *Strictly English: The correct way to write … and why it matters* (2010) includes the following entry on data: “Certain words, usually of foreign origin, are not always recognised by people as being plural. *Data* and *media* are the plurals of anglicised Latin neuter nouns and should take verbs as such – “the data were wrong” or “the media are scum”” (p. 53).

[2] A full description of the Stenton Corpus can be found in Stenton (in progress).

[3] We are very grateful to Cambridge University Press for permission to hold the files that comprise the Stenton Corpus. The six journals used for this study are: *Asian Journal of International Law* (AJL), *Asian Journal of Law and Society* (ALS), *Bilingualism: Language and Cognition* (BLC), *International Journal of Law in Context* (IJC), *Journal of Child Language* (JCL), and *Language and Cognition* (LCO).

[4] Mauranen (2012) and Crystal (2017, p. 206) prefer “English as a Lingua Franca”, but both of these are writing in the context of spoken academic English.

[5] Potential revisions shown here in brackets were not included in the survey extracts.

References

APA (2020 seventh edition) *Publication manual of the American Psychological Association*. American Psychological Association.

Cameron, D. (2012). *Verbal hygiene* (2nd edition). Routledge.

Crystal, D. (2017) *Making sense: the glamorous story of English grammar*. Profile Books.

Curzan, A. (2014). *Fixing English: prescriptivism and language history*. Cambridge University Press.

Davies, M. (2008). *The corpus of contemporary American English (COCA)*. http://corpus.byu.edu/coca/ [last accessed 11 August 2021].

Davies, M. (2013). *Corpus of global web-based English (GloWbE)*. http://corpus.byu.edu/glowbe/ [last accessed 11 August 2021].

Ebner, C. (2014). The dangling participle – a language myth? *English Today*, *30*(4), 3–4. https://doi.org/10.1017/S0266078414000327

Ebner, C. (2017). *Proper English usage: a sociolinguistic investigation of attitudes towards usage problems in British English*. LOT.

Ebner, C. (2018). Attitudes to British usage. In I. Tieken-Boon van Ostade (Ed.), *English usage guides: history, advice, attitudes* (pp. 137–154). Oxford University Press.

Einsohn, A. (2006). *The copyeditor's handbook: a guide for book publishing and corporate communications*. University of California Press.

Gilman, E. W. (Ed.). (1989). *Webster's dictionary of English usage*. Merriam-Webster.

Heffer, S. (2010). *Strictly English: the correct way to write … and why it matters*. Random House Books.

Kostadinova, V. (2018). *Language prescriptivism: attitudes to usage vs. actual usage in American English* [Doctoral dissertation, University of Leiden]. Leiden University Scholarly Publications. https://hdl.handle.net/1887/68226

Lukač, M. (2018). *Grassroots Prescriptivism*. LOT.

Mauranen, A. (2012). *Exploring ELF: Academic English shaped by non-native speakers*. Cambridge University Press.

McArthur, T. (2001). Error, editing, and World Standard English. *English Today*, *17*(1), pp. 3–8. https://doi.org/10.1017/S0266078401001018

McEenery, T., & Hardie, A. (2012). *Corpus linguistics: method, theory and practice*. Cambridge University Press.

Milroy, J., & Milroy, L. (2012) *Authority in language: investigating language prescription and standardisation* (4th ed.). Routledge.

Mittins, W. H., Salu, M., Edminson M., & Coyne, S. (1970). *Attitudes to English usage: an enquiry by the University of Newcastle upon Tyne Institute of Education English Research Group*. Oxford University Press.

Owen, J. (2020). Practicing prescriptivism: how copyeditors treat prescriptive rules. In D. Chapman & J. D. Rawlins (Eds.), *Language prescription: values, ideologies and identity* (pp. 292–306). Multilingual Matters.

Peters, P. (2004). *The Cambridge guide to English usage*. Cambridge University Press.

Peters, P. (2018). The lexicography of English usage. In I. Tieken Boon van Ostade (Ed.), *English usage guides: history, advice, attitudes* (pp. 31–49). Oxford University Press.

Pillière, L. (2018). Imposing a norm: the invisible marks of copy-editors. In L. Pillière, W. Andrieu, V. Kerfelec, & D. Lewis (Eds.), *Standardising English: norms and margins in the history of the English language* (pp. 251–276). Cambridge University Press.

Pillière, L. (2020). US copy editors, style guides and usage guides and their impact on British novels. In D. Chapman & J. D. Rawlins (Eds.), *Language prescription: values, ideologies and identity* (pp. 261–291). Multilingual Matters.

Rawlins, J. D. & Chapman, D. (2020). Introduction: values and binaries in language evaluation. In D. Chapman & J. D. Rawlins (Eds.), *Language prescription: values, ideologies and identity* (pp. 1–11). Multilingual Matters.

Rogers, S. (2012, June 8). Data are or data is? *The Guardian*. www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular

Sayce, K. (2006). *What not to write: a guide to the dos and don'ts of good English*. Words at Work.

Severin, A. A., & Burridge, K. (2020). What do 'Little Aussie Sticklers' value most? In D. Chapman & J. D. Rawlins (Eds), *Language Prescription: values, ideologies and identity* (pp. 194–211). Multilingual Matters.

Stenton, A. (in progress). *These Kind of Words: Number agreement in the species noun phrase in International Academic English* [Doctoral dissertation, Leiden University Centre for Linguistics].

Straaijer, R. (2014). *Hyper usage guide of English*. http://huge.ullet.net

Taggart, C. (2010) *Her Ladyship's guide to the Queen's English*. National Trust.

Tieken-Boon van Ostade, I. (2013). Studying attitudes to English usage. *English Today*, *29*(4), 3–12. https://doi.org/10.1017/S0266078413000436

Tieken-Boon van Ostade, I. (2020). *Describing prescriptivism: usage guides and usage problems in British and American English*. Routledge.

Tieken-Boon van Ostade, I., & Ebner, C. (2017). Prescriptive attitudes to English usage. *Oxford research encyclopedia: Linguistics*.

Trudgill, P., & Hannah, J. (2017) *International English: a guide to varieties of English around the world* (6th ed.). Routledge.