DE GRUYTER
OPEN

ICAME Journal, Volume 42, 2018, DOI: 10.1515/icame-2018-0005

Charting orthographical reliability in a corpus of English historical letters

Anni Sairio, Samuli Kaislaniemi, Anna Merikallio and Terttu Nevalainen
University of Helsinki[1]

Abstract

Research into orthography in the history of English is not a simple venture. The history of English spelling is primarily based on printed texts, which fail to capture the range of variation inherent in the language; many manuscript phenomena are simply not found in printed texts. Manuscript-based corpora would be the ideal research data, but as this is resource-intensive, linguists use editions that have been produced by non-linguists. Many editions claim to retain original spellings, but in practice text is always normalized at the graph level and possibly more so. This does not preclude using such a corpus for orthographical research, but there has been no systematic way to determine the philological reliability of an edited text. In this paper we present a typological methodology we are developing for the evaluation of orthographical quality of edition-based corpora, with the aim of making the best use of bad data in the context of editions and manuscript practices. As a case study, we apply this methodology to the Early Modern and Late Modern English sections of the Corpus of Early English Correspondence.

1 Introduction: Challenges for studying historical manuscript spelling practices

The study of orthography in the history of English is not a simple venture. Although EEBO-TCP now provides us with one and a half billion words of printed Early Modern English, many orthographical features are not found in printed texts, and access to private spellings as recorded in manuscripts is impeded by the lack of suitable resources of any scale. On the one hand, corpora such as the *Corpus of Early English Correspondence* (CEEC) and ARCHER have been based on edited texts and therefore cannot be taken to represent manuscript reality. There is no denying the "pervasive, if usually covert" (Fulk

2017: 434) role of philology in historical linguistics, given that "almost without exception, such studies rely on edited texts [...] and they interpret the linguistic components of texts on the basis of tools compiled by philological methods" (2017: 434). It is thus unclear whose language we are actually seeing in edited texts. On the other hand, manuscript-based resources edited by philologists, such as the *Electronic Text Edition of Depositions 1560–1760* (ETED), are much restricted in size and scope because editing primary texts demands considerable resources. Previous work on Early Modern English private spelling practices has been hampered by both of these restrictions, and the results of such studies are either long-term and superficial (Osselton 1984), or focussed on a specific time, place, or text (Sönmez 1993). Overall, our understanding of the history of English spelling is largely based on printed sources (Scragg 1974; Salmon 1999).

Paradoxically, while modern editions of historical manuscript texts ostensibly allow us to study the language of manuscripts, they are the very obstacle to accessing historical private spelling practices. Although many editions claim to retain original spellings, in practice editors always normalize texts at the graph level, and some practices of silent editorial normalization have become so conventionalized that they may not even be mentioned in the editorial principles. For instance, editions of Early Modern English texts as a rule modernize punctuation, capitalization, and <u/v>-variation (changing spellings like *giue, vp* to *give, up*). Other frequently silent editorial practices include the modernization of word division (*to morrow* > *tomorrow*, *your self* > *yourself*), and the expansion of ampersands to *and*. Editorial modernization and normalization is, of course, not restricted to historical English, but an issue irrespective of language or time period. The nature of problems also changes over time, as the temporal / linguistic distance between a text and its editor increases and the editor must carefully consider factors of readability and accessibility.

In the end, we simply do not know to what extent even meticulous editors have quietly changed the text, not to mention other manuscript features such as layout. And the problem of such known editorial interference is compounded by editorial errors. For example, Grund, Kytö and Rissanen (2004: 147–148) show how the Boyer and Nissenbaum (1977) edition of the Salem Witchcraft records contains misrepresentations and omissions, and the CEEC team have compared several good-quality editions against manuscripts and discovered inconsistencies (see Section 2).

It is therefore almost a truism that editions do not provide the best data for the investigation of private spelling, and it would no doubt be every scholar's preference to use large manuscript-based corpora for spelling research. However, the suitability of edited data for orthographical research has not been seriously and systematically investigated, and this is the aim of the ERRATAS project and the present paper. How can we make the best use of bad data (see e.g. Nevalainen 1999) in the research of historical orthography? Can we make an informed decision to select specific editions for the study of manuscript spelling? This paper introduces the methodology we are developing to explore these questions.

As a case study, we chart the philological reliability of the seventeenth- and eighteenth-century sections of CEEC. There is no exhaustive list of orthographical features that we can apply: by doing this research we are also documenting and mapping orthographical reality in the Early and Late Modern English periods. Many of the editions in CEEC are by large publishers such as Cambridge University Press and Oxford University Press, and they are also included in online resources like *Early Modern Letters Online* and *Electronic Enlightenment*, so this analysis has also broader relevance with regard to digital databases.[2]

The structure of this paper is as follows. In Section 2, we introduce CEEC and discuss editorial principles and complexities. In Section 3, we present an overview of the methodology we have developed to assess the philological reliability of the corpus collections. In Section 4, we present our findings on the seventeenth-century section and what we can determine of the eighteenth-century section at this time. Section 5 provides an overview of our main aims and current conclusions.

2 The CEEC and editorial principles

The *Corpus of Early English Correspondence* (CEEC) is a single-genre corpus of personal letters that spans from 1400 to 1800 (www.helsinki.fi/varieng/CoRD/corpora/CEEC/; Raumolin-Brunberg and Nevalainen 2007; Nevala and Nurmi 2013). It consists of 5.3 million words from 12,000 letters in 189 editions. CEEC does not include entire editions, and the selections of letters from the editions are referred to as collections. CEEC was originally designed for the sociolinguistic study of morphology, but the linguistic phenomena that have been studied using this corpus now include syntax, pragmatic phraseology, and grammaticalized lexemes.

The corpus compilers have maintained that the CEEC corpus family “cannot provide reliable evidence for the study of e.g. orthography, punctuation or visual prosody”, given that “[t]hese features are inadequately represented in most printed editions” (Nevala and Nurmi 2013: section 2.3). As Nevala and Nurmi (2013: section 2.3) note, some of the editions indeed “normalise many details of interest for linguists, such as abbreviations, superscripts, spelling variation and the like". The CEEC team has attempted to tackle this issue by checking editions against manuscript letters, by editing some letters from manuscripts, such as the Gawdy letters by Minna Nevala, and by re-editing poor collections (Keränen 1998). Overall, the scholars working on CEEC have avoided spelling research (but see Kaislaniemi et al. 2017 for a longitudinal study that compares spelling features in CEEC with manuscript evidence).

In the corpus compilation process, only those letters were selected which the editor had transcribed from the existing originals. Spelling was a key selection criterion in the compilation process of CEEC, and the aim of the compilers was always to include 'original-spelling' editions. 'Original spelling' was evaluated on the basis of what the editors state as their principles and what the letters themselves indicate. A closer look at editions and manuscripts nevertheless reveals that editorial practices and the degree of meticulousness vary within individual editions. The editor may declare to have reproduced the spelling of the original manuscripts, but may still deviate from this principle. *The Letters of Joseph Addison* are stated to be published "in as complete as form as possible, and precisely as written as when the originals are still in existence" (Graham 1941, preface). Comparison with Addison's manuscript letters in the British Library shows that the editor has retained high-frequency contractions and abbreviations (*confess'd, y^{e}*), almost all of Addison's capitalization (*Late Tumults, Kingdome, Retainer*), and superscripts (*y^{e}, Dec^{r})* in the test letters. On the other hand, the editor has not been systematic in their practices, retaining as well as changing spelling features. Affected features include capitalization, textual deletions and insertions, verb contractions (*woud* > *would, allowd* > *allowed*), vowel variation (*Rhime* > *Rhyme*) and present-day English compound words (*any thing* > *anything, can not* > *cannot*). The letters are thus not published "precisely as written", as the editor states in the preface.

Nevertheless, Graham (1941) is a high-quality edition for the purposes of CEEC (see also Kaislaniemi 2017: 52–59), and in terms of spelling, considerably more has been retained than changed. The edition can be used for the study of select orthographical features in the eighteenth century. While Graham (1941) would not provide reliable information of the *-d* contractions of preterites and past participles, nor of modal contractions, it could be used in a study of extra initial capitalization (with caveats), and possibly to examine the use of apostrophised contractions (bearing in mind that the variant form *-d* does not appear to be retained). Analysing the philological reliability of even a single edition is therefore a complex matter.

It should be noted that the issues pertaining to editing historical texts also apply to corpus compilation. Because the media are different (printed book vs. electronic text), and the purposes of the text are different (reading vs. corpus queries), not all textual features in editions are preserved in corpora. Thus, where an edition may change a feature, such as abbreviation markers, for typographical reasons, CEEC, being ASCII text, changes superscripts into tags: for example, *y*^{e} becomes `y=e=` (see Nurmi 1998, ch. 2 for CEEC coding conventions).

In the next section, we present our method of philological inquiry.

3 Methods: How to approach private spelling practices in edition-based corpora

The ERRATAS project has two goals, both built on the assessment of the philological reliability of printed editions of historical letters. Such assessments will allow us to identify 1) not only those editions used in CEEC that can be used to study spelling, but also 2) the specific textual linguistic features that can be investigated using CEEC. We also hope to create a typology or a scale to rank editions according to their philological reliability. Such a typology would be applicable to any edition of Early Modern English material, and adaptable to other periods and languages.

In this process, we comb through the corpus collections and their source editions with a **checklist** that **charts textual features**, in order to **determine orthographical reliability** and the degree of editorial interference. The information is fed into an MS Access **database**. This process enables us to identify the corpus collections which can be used for particular types of spelling research. The next stage will be to study the data to create a **typology of editorial reliability**, which can then be used to **tag corpus collections**, allowing users of a corpus to **fine-tune their corpus queries** according to the level of philological rigour necessary for their research questions. A tangential step in the process will be to **check the source manuscripts** when necessary, in order to determine the veracity of our findings. The process of investigation is illustrated in Figure 1:

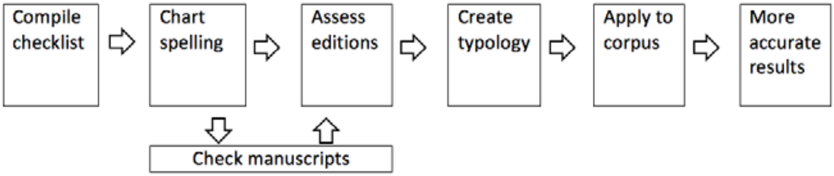

Figure 1: ERRATAS workflow

We have refined the method through a series of iterative rounds.

English spelling only began to undergo standardization from the first half of the seventeenth century. This means that in seventeenth-century manuscript texts, we expect to see a wide range of spelling variation, including allographic variation of <u> and <v>, the use of <i> for /j/ and the gradual adoption of <j>, variation in word-medial vowel clusters (e.g. *beleve, beleeve, beleive, believe*), and the doubling or singling of consonants (e.g. *att* for *at*, *al* for *all*). Manuscript texts also usually contain textual features not found in print, such as certain types of abbreviations and the liberal use of superscripts. Therefore, it is in principle possible to establish the degree of editorial 'interference' in an edition of a seventeenth-century manuscript text, based on how many textual features that we expect to see are indeed present.

Testing this idea is one of the aims of the ERRATAS project. It is of course obvious that any conclusions drawn by this method are not conclusive: only the comparison of an edited text with the source manuscript will reveal the full editorial fingerprint. To control for this, the ERRATAS project will also check editions used in CEEC against their source manuscripts. But if we can conclude that the philological reliability of an edition can be determined without recourse to the source manuscripts – even if for just some aspects of the text – we can dramatically increase the amount of edited texts that can be used for historical linguistics on the one hand, and the scope of what can be investigated in them on the other.

The keystone of the ERRATAS system is a checklist of textual features (see below) to be surveyed in editions: they consist of features we can expect to find in historical English texts. Given the lack of baseline data, the checklist also contains features that can only begin to be investigated with large-scale data. The system was initially designed for seventeenth-century texts. It was expanded and updated for the eighteenth-century analysis.

The ERRATAS checklist gives detailed information on how each textual feature can be charted in the editions and corpus texts, suggests audit words for concordance searches, and, based on our current understanding, gives a rough estimate for when each feature has been in use. The information retrieved on the basis of the checklist is recorded in an Access database via two separate Access forms: one for the editorial conventions (what the editor explicitly states to have done), and another one for features charted in the edited text. This makes it easier to determine whether the variation or the lack thereof is an editorial or an authorial decision.

The checklist, and thus the database data feed form, currently contains 230 data fields, several for each of the c.100 textual features (see Appendices 1 and 2). Textual features are collected into 13 categories:

1. Text (e.g. presence of opening formulas, letter superscriptions)
2. Allographs (e.g. the 'long s')
3. Spelling (e.g. *give*/*giue* or *enjoy*/*enioy,* -ED variation)
4. Capitalization (e.g. *Late Tumults*)
5. Special characters and obsolete graphs (e.g. the thorn <þ> for <th>)
6. Macrons and abbreviation markers (e.g. *cõmon*/*common*)
7. Superscripts (e.g. w^{ch}/*which*, fo^{r}/*for*)
8. Word division (e.g. *to morrow* / *tomorrow)*
9. Morphology (e.g. *more than* / *more then*)
10. Abbreviations (e.g. y^{t}/*that, dd/deliver*)
11. Punctuation (e.g. virgules, commas, full stops, apostrophes)
12. Layout (e.g. text in margins, blank spaces in running text)
13. Extratextual information (e.g. retained misspellings or deletions)

Some textual phenomena contain several features, in as many categories. For instance, an Early Modern English spelling like m^{r}*chãt* 'merchant' contains a macron (indicating a missing <n>), a superscript (standing for <er> or <ar>), and of course is itself an abbreviation, and would thus be recorded under categories 6, 7 and 10. While the presence of abbreviation markers would correlate with the presence of abbreviations, the reverse is not true (cf. abbreviations like *pd* 'paid'); and the same applies for superscripts, as they can be otiose (e.g. fo^{r}).

We found that it was necessary to have a threshold of occurrences before recording a textual feature as present or absent in a text. Editors make mistakes, and even a normalized-spelling edition may retain scattered instances of 'old form' spellings. Yet such editions must nonetheless be categorized as having normalized spellings, and this is recorded in the ERRATAS database accordingly: but a note is added to record the presence of any outliers. Some features cannot be determined by reference to the edition alone, unless the editor indicates their practices in detail. These include the 'long s', some obsolete graphs, hyphens, and spacing. We have a database field for 'applicability', which charts whether it is at all possible to determine if a feature is present in the source text, based on the edition alone. These features will be confirmed by consulting the source manuscripts.

Formulating this investigation into a binary, machine-readable yes/no procedure without leaving out any important information is not simple. For example, if a feature does not appear in the text, it does not necessarily mean that the editor has not included it. Lack of data may result from damage done to the manuscript, or from authorial decisions in e.g. avoiding variation in the spelling of names. The database thus includes important notes on how to interpret the content, and not all findings can be simplified into numerical data.

Overall, charting spelling variation in a corpus is not a straightforward process. Simple concordance searches of keywords (*relieve/releive*, *friend/freind*) usually reveal whether there is variation. If an edition does not contain notable spelling variation overall, there is a chance that e.g. the category of names will reveal variant spellings. However, names have not been tagged in CEEC, and that leads to more time-consuming corpus searches. The iterative nature of the process is emphasized; in order to make an informed decision about the presence or absence of a feature, the researcher needs to go back and forth between the editions and the corpus texts since, as noted above, the corpus does not contain entire editions, nor retain all the textual features of the printed text.

In sum, the ERRATAS method is to go through editions (and corpus texts) following the purpose-built checklist of textual features, and record findings in a database. This data can then be used to determine which textual features in an edition/corpus are likely to be philologically reliable – and, after consulting the original manuscripts, which textual features are verified to be so. In the next section we present our preliminary findings for CEEC. The dataset for the first analysis consists of 34 editions of seventeenth-century letters, and for the second analysis we surveyed 65 editions of eighteenth-century letters. (For a list of the editions surveyed, see Appendix 3).

4 Orthographical evaluation of CEEC: Preliminary results

Editions of seventeenth-century letters allow us to draw some conclusions about their philological reliability without comparisons with the manuscript data. But due to the coexisting use of what we now consider as standard spelling and older styles of spelling, eighteenth-century orthography is more complex and requires access to the manuscript data before the editorial fingerprint can be confirmed.

4.1 Editions of seventeenth-century letters

Our analysis of seventeenth-century letter collections in CEEC proves that the ERRATAS system of identifying reliable research material works. The process turned out to be more time-consuming than expected, however, and the initial analysis covers only 34 editions (879,820 words in CEEC; the entire seventeenth-century CEEC comprises 1,931,205 words). This section reports three initial findings from this material.

The first preliminary study of the results looks at <u/v>-variation. Specifically, the following strings were searched in the corpus: *give**, *giue**, *geve**, *geue**; *have**, *haue**; *ever*, *euer*; *up**, *vp**; *un-**, and *vn-**. The results, shown in Figure 2, indicate that the amount of <u/v>-variation increases considerably with the application of the ERRATAS method and the identification of philologically more reliable editions, which ought to be used for this type of research. Initially, in all the corpus texts drawn from these 34 editions, 'new form' spellings (i.e., spellings that became standardized: *give, have, ever, up*, un**) dominate with 87 per cent of the tokens. Using the ERRATAS method, 15 out of these 34 editions can be identified as retaining <u/v>-variation – and, indeed, other original-spelling features. In this new subcorpus (332,047 words), which can thus be considered to better represent seventeenth-century manuscript reality than the 34 editions together, we can see that <u/v>-variation increases to 31 per cent, and 'new form' spellings comprise only 69 per cent of the tokens.[3]

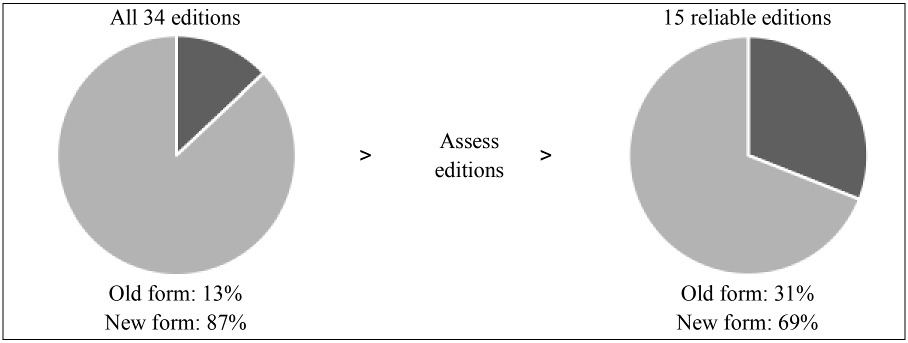

Figure 2: <u/v>-variation in seventeenth-century letters: give, giue; up, vp

This difference is striking. In terms of the stages of linguistic change, what, based on the initial evidence, appeared to be spelling standardization that was in effect completed, is shown to have only just reached the previous stage of nearing completion (see Nevalainen and Raumolin-Brunberg 2016: 54–55). Although 'new form' spellings were certainly vibrant at the beginning of the century – indeed, some writers used them consistently, for some words – both old and new spellings were standard in the sense that they were allowed, and no stigma was attached to either. Variation between old and new spellings was the norm, rather than the exception, until the end of the century. The unsorted data gives the impression that spelling standardization was further along in handwritten texts than it actually was: this view is corrected by the application of the ERRATAS method, which provides a more reliable starting point for orthographical analysis.

A final caveat is in order: there is currently no data regarding the expected ratio of old to new spellings in manuscript texts in the Early or Late Modern English period. Expanding this preliminary study to cover all <u> and <v> spellings in English would certainly increase the accuracy of the figures, but without recourse to the manuscripts, we are ultimately unable to establish conclusively whether such results are valid. One way of correcting for this inaccuracy is to compare these figures to data from resources or corpora formed of philologically rigorous transcriptions of manuscript texts. Even allowing for variation across time and space and writer, this should allow us to establish the veracity of results gained through the application of the ERRATAS method.

The second pilot study based on ERRATAS data for the seventeenth-century parts of CEEC is a 'collocation analysis' of editorial changes. One research question of the ERRATAS project is whether editorial practices cluster. That is, if editors change feature X, do they also change feature Y, or Z? This study charted approximately 100 different textual features in 41 editions, and Table 1 lists the proportional presence of 22 of these features. This list suggests that editorial changes fall on a scale, instead of emerging as distinct clusters.

Table 1: 22 of the c.100 textual features charted in CEEC; proportional presence of editorial changes in 41 editions of seventeenth-century letters

100%	*-es* graphs changed Line breaks changed
80–89%	Capitalization partly normalized Virgules changed Majuscule <I/J> modernized p-graphs changed
68–76%	Blanks changed <ff/F> modernized Macrons and abbreviation markers omitted
50–59%	<u/v> modernized Otiose superscripts lowered (y^e, fo^r)

36–44%	y-as-thorn (y^e) changed Abbreviations changed Quotation marks added Superscripts lowered Superscripts for dates, numbers and money lowered Minuscule <i/j> modernized
17–29%	Scribal emendations marked Apostrophes added Word division normalized
5–10%	Names normalized Capitalization modernized

Table 1 tells us, for example, that while line breaks are essentially never retained in editions, names are only rarely normalized. The presence of less frequent types of editorial changes suggests that more common changes have also been made – however, whether this is the case remains to be charted. Note, however, that editors do not treat related features in the same way: although capitalization is rarely completely modernized, it is almost always partly normalized (usually this applies to names and sentence-initial capitals in particular).

The third initial study of the seventeenth-century parts of CEEC addresses the question of changes in editorial practices over time. Figure 3 looks at 22 textual features in editions that date from a period spanning two centuries. As can be seen, older editions do not seem to be necessarily worse in terms of philological rigour. The same appears to apply to the eighteenth-century materials, as can be seen below in Figure 4.

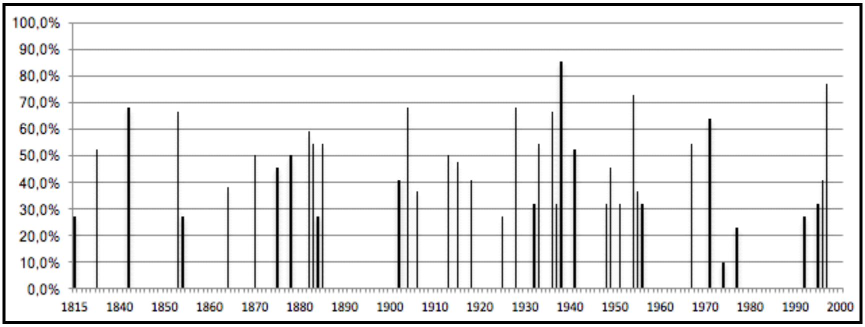

Figure 3: Quality of edition vs. publication date: 22 textual features in 41 editions of seventeenth-century letters, percentage of 'old form' retained (high is good)

4.2 Editions of eighteenth-century letters

Standardization of private spelling in English had not reached conclusion by the end of the eighteenth century, and in this period, we see modern orthographical styles, preferences of outgoing 'non-standard' styles, and mixtures of both. As there is no large-scale baseline data, we cannot state which spelling features we would expect to see in eighteenth-century letters. Unlike in the seventeenth-century material, the absence of certain 'non-standard' forms does not necessarily signal modernization by the editor. This is a diverse century in terms of how standardization proceeds in time and through the social strata, and educated letter-writers of the early eighteenth century seem to vary their spelling in many ways more than educated writers at the end of the century (see e.g. Oldireva Gustafsson 2002 for generational differences in letterwriting, though in edited data).

Without the availability of original manuscripts, we present only the first step of our work. At this point, we have charted the presence of variation in 54 textual features in the 65 editions of the eighteenth-century part of CEEC (CEECE), 19 of which are shown in Table 2.

Table 2: 19 of 54 orthographical features charted in CEECE; proportional presence of variation in textual features

100%	Abbreviations Apostrophes
88–97%	Name variation <ie/ei> variation -ED variation Variation in capitalization Dashes Long vowel and diphthong variation Variation in word division Doubled consonants Ampersand
48–60%	Variation in sentence capitalization Superscripts Superscripts for dates, numbers and money y-as-thorn (y^e)
17–23%	Otiose superscripts Separate prefixes Macrons and abbreviation markers Variation in name capitalization

Where Table 1 presented the results of our analysis regarding editorial changes, Table 2 only records the presence of variation in the editions, and at this point, Table 2 cannot be used to draw further conclusions about editorial practices. For example, the absence of ampersands may result from editorial decisions, but it is also possible that the editions are faithful to the original manuscripts. We also know that *y*^{e} as a definite article occurs in 48 per cent of the editions, but we do not know (and cannot tell from this data) whether *y*^{e} has by and large been replaced in the other editions, or whether *y*^{e} simply does not occur very often in eighteenth-century manuscript sources. In the Bluestocking Corpus (unpublished version, twelve writers, 1730s–1790s), the frequency of *y*^{e} for *the* is 17 per cent, and not all of the letter-writers in the corpus use this variant.

That said, the data does suggest that the following features are common in eighteenth-century manuscripts, and have been widely retained in editions of the same (percentages of editions that include these features):

97% name variation (*Bowry/Bowery/Bowrey*)

97% <ie/ei> variation (e.g. *beleive*)

97% -ED variation (e.g. *receivd*)

95% variation in capitalization (*Your Letter occasioned in Me a Fit of Passion*)

94% long vowel and diphthong variation (*agreable, bin/been*)

94% variation in word division (*to morrow, him self*)

91% doubled and singled consonants (*att, maried*)

97 per cent of the CEECE editions thus contain name variation, word-medial <ie/ei> variation, and preterite and past participle -ED variation. Our next step will be to compare the edited letters with manuscript images when possible; we anticipate being able to do so with a number of collections. We will also chart this variation with more fine-grained diachrony, observing whether e.g. -ED variation occurs consistently throughout the century.

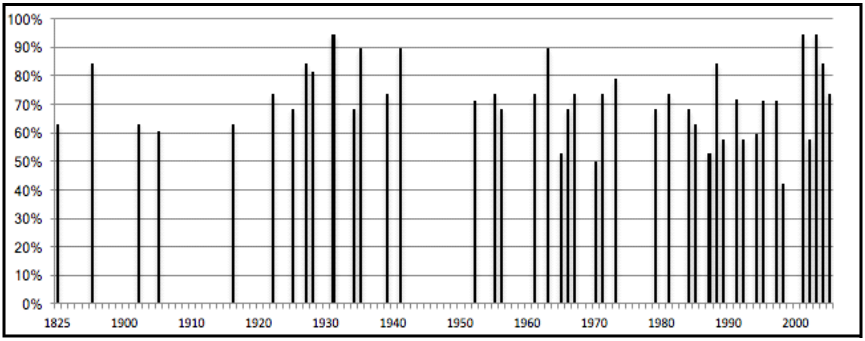

Figure 4: Quality of edition vs. publication date: 19 textual features in 65 editions of eighteenth-century letters, percentage of presence in edition (high is good)

Present-day editions are again not necessarily more accurate than the older editions (Figure 4; cf. Figure 3), if we assume that the more variation in textual features we see, the more the editions resemble manuscript reality. Of interest is also the diachronic makeup of the corpus. The majority of the CEECE editions are from the twentieth century, with only two from the nineteenth century, whereas the seventeenth-century corpus section contains thirteen editions from the nineteenth century. Quite a few nineteenth-century editions did not pass the inspection when CEECE was compiled; many nineteenth-century editors were family members intent on preserving a specific image of their eighteenth-century ancestors, a process which often involves omitting sensitive topics and polishing the perceived irregularities of spelling and style.

5 Conclusion

Historical corpora of English are primarily compiled from editions. The results of such ‘philological outsourcing’ has enabled research on e.g. syntax and pragmatics, but linguists have been reasonably cautious about studying orthography using edition-based corpora, even when the source editions claim to retain original manuscript spelling. This compromise has been accepted in order to create e.g. the 5.3-million-word CEEC. The use of editions does not preclude using a corpus for orthographical research, but there has been no systematic way to determine the philological reliability of the edited texts and, thus, of the corpus. And given the likelihood of silent standardisation and normalization in editions, we can ask: whose language are we seeing?

This paper has presented a tool that we are developing in order to assess the orthographical quality of edition-based corpora. By investigating how editors have reproduced the original spelling in the manuscripts, we seek to identify the scope and possibilities of CEEC for longitudinal orthographical research. Our aim is to extend the principle of 'making the best use of bad data' (e.g. Nevalainen 1999) to the field of manuscript practices and edited data. This will make edited material accessible for orthographical analysis in a more systematic way, and allow for new openings in historical linguistic research.

In this pilot study, we have evaluated altogether 99 collections of seventeenth- and eighteenth-century letters with a checklist of textual features that we would expect to see in letters of the period. One end goal is to assign editions – and, thus, corpus texts – a rating of orthographical reliability. Certain features are expected to be present in seventeenth-century manuscript texts, and their absence signals the editor's decision to omit them, making it possible to tentatively evaluate the reliability of an edition. In contrast, eighteenth-century data requires manuscript evidence to confirm what the presence or absence of spelling features in the editions means. Manuscript evidence shows that individual Late Modern period writers vary between modern orthographical styles, preferences for outgoing 'non-standard' styles, and mixtures of both (e.g. Kaislaniemi et al. 2017). Our next step will be to compare the eighteenth-century corpus collections against manuscript images recently photographed in archives.

We have discovered that the phrase 'original spelling retained' has meant different things to different editors (and publishers). At this point we are able to conclude that CEEC can indeed be used to study highly common spelling features (cf. Kaislaniemi et al. 2017). However, this assessment comes with caveats and reservations, and the work is ongoing. The initial results from the ERRATAS project in turn show that the method can be fruitfully used to determine the philological reliability of editions of historical texts. Thus it is, after all, possible to use edition-based corpora to study textual features previously presumed distorted or inaccessible.

Notes

1. The authors would like to thank Dr Tanja Säily for her comments that greatly improved the final article.
2. ERRATAS is part of the larger research project Interfacing structured and unstructured data in sociolinguistic research on language change (STRATAS, 2016-2019, Academy of Finland, blogs.helsinki.fi/stratas-project/).

3. In the 19 editions that do not retain <u/v>-variation, the new form is used 97% of the time. This figure is not 100% due to editorial inconsistencies and errors, and also possible errors committed in the analysis: the results presented here are work in progress.

References

The Bluestocking Corpus: Private Correspondence of Elizabeth Montagu, 1730s–1780s. First version. Edited by Anni Sairio, XML encoding by Ville Marttila. Department of Modern Languages, University of Helsinki. 2017. bluestocking.ling.helsinki.fi.

Boyer, Paul and Stephen Nissenbaum. 1977. *The Salem witchcraft papers: Verbatim transcripts of the legal documents of Salem witchcraft outbreak of 1692*. 3 vols. New York: Da Capo Press.

Early Modern Letters Online (EMLO). Cultures of Knowledge, Bodleian Libraries, University of Oxford. emlo.bodleian.ox.ac.uk.

Electronic Enlightenment. Oxford: Bodleian Libraries, University of Oxford. www.e-enlightenment.com.

An Electronic Text Edition of Depositions 1560–1760 (ETED). 2011. Compiled by Merja Kytö, Peter J. Grund and Terry Walker. Available on the CD accompanying Merja Kytö, Peter J. Grund and Terry Walker (eds.), *Testifying to language and life in Early Modern England*. Amsterdam/Philadelphia: John Benjamins.

Fulk, Robert D. 2017. Philological coda. Noise: An appreciation. *English Language and Linguistics* 21 (2): 431–438.

Graham, Walter (ed.). 1941. *The letters of Joseph Addison*. Oxford: Clarendon Press.

Grund, Peter, Merja Kytö and Matti Rissanen. 2004. Editing the Salem Witchcraft records: An exploration of a linguistic treasury. *American Speech* 79 (2): 146–166.

Kaislaniemi, Samuli. 2017. Reconstructing merchant multilingualism: Lexical studies of early English East India Company correspondence. PhD thesis, University of Helsinki.

Kaislaniemi, Samuli, Mel Evans, Teo Juvonen and Anni Sairio. 2017. 'A graphic system which leads its own linguistic life'? Epistolary spelling in English, 1400–1800. In T. Säily, A. Nurmi, M. Palander-Collin and A. Auer (eds.). *Exploring future paths for historical sociolinguistics* (Advances in Historical Sociolinguistics 7), 187–214. Amsterdam: John Benjamins.

Keränen, Jukka. 1998. Forgeries and one-eyed bulls: Editorial questions in corpus work. *Neuphilologische Mitteilungen* 99 (2): 217–226.

Nevala, Minna and Arja Nurmi. 2013. The Corpora of Early English Correspondence (CEEC400). In A. Meurman-Solin and J. Tyrkkö (eds.). *Principles and practices for the digital editing and annotation of diachronic data* (Studies in Variation, Contacts and Change in English 14). Helsinki: VARIENG. www.helsinki.fi/varieng/series/volumes/14/nevala_nurmi/.

Nevalainen, Terttu and Helena Raumolin-Brunberg. 2016. *Historical sociolinguistics: Language change in Tudor and Stuart England*. 2nd ed. New York: Routledge.

Nevalainen, Terttu. 1999. Making the best use of 'bad' data: Evidence for sociolinguistic variation in Early Modern English. *Neuphilologische Mitteilungen* 100 (4): 499–533.

Nurmi, Arja (ed.). 1998. *Manual for the Corpus of Early English Correspondence Sampler CEECS*. Helsinki: Department of English, University of Helsinki. Available at clu.uni.no/icame/manuals/CEECS/INDEX.HTM.

Oldireva Gustafsson, Larisa. 2002. *Preterite and past participle forms in English 1680–1790: Standardisation processes in public and private writing*. Uppsala: Acta Universitatis Upsaliensis.

Osselton, Noel. 1984. Informal spelling styles in Early Modern English: 1500–1800. In N.F. Blake and C. Jones (eds.). *English historical linguistics. Studies in development*, 123–137. Sheffield: Department of English Language, University of Sheffield.

Raumolin-Brunberg, Helena and Terttu Nevalainen. 2007. Historical sociolinguistics: The *Corpus of Early English Correspondence*. In J.C. Beal, K.P. Corrigan and H.L. Moisl (eds.). *Creating and digitizing language corpora*. Vol. 2: *Diachronic databases*, 148–171. Houndsmills: Palgrave Macmillan. Pre-print available at www.helsinki.fi/varieng/CoRD/corpora/CEEC/generalintro.html.

Salmon, Vivian. 1999. Orthography and punctuation 1476–1776. In R. Lass (ed.). *The Cambridge history of the English language. Volume III: 1476–1776*, 13–55. Cambridge: Cambridge University Press.

Scragg, Donald G. 1974. *A history of English spelling*. Manchester: Manchester University Press.

Sönmez, Margaret J.-M. 1993. English spelling in the seventeenth century: A study of the nature of standardisation as seen through the MS and printed versions of the Duke of Newcastle's *A New Method*. PhD Thesis, University of Durham.

Walker, Terry and Merja Kytö. 2013. Features of layout and other visual effects in the source manuscripts of *An Electronic Text Edition of Depositions 1560–1760* (ETED). In A. Meurman-Solin and J. Tyrkkö (eds.). *Principles and practices for the digital editing and annotation of diachronic data* (Studies in Variation, Contacts and Change in English 14). Helsinki: VARIENG. www.helsinki.fi/varieng/series/volumes/14/walker_kyto/

Appendix 1: ERRATAS checklist of editorial interference with audit word lists

doi.org/10.5281/zenodo.1187074

Appendix 2: ERRATAS database data entry form

doi.org/10.5281/zenodo.1187076

Appendix 3: List of editions surveyed for ERRATAS

doi.org/10.5281/zenodo.1187078