DE GRUYTER OPEN

ICAME Journal, Volume 39, 2015, DOI: 10.1515/icame-2015-0001

Guidelines for normalising Early Modern English corpora: Decisions and justifications

Dawn Archer[1], Merja Kytö[2], Alistair Baron[3] and Paul Rayson[3]
Universities of Central Lancashire[1], Uppsala[2] and Lancaster[3]

Abstract

Corpora of Early Modern English have been collected and released for research for a number of years. With large scale digitisation activities gathering pace in the last decade, much more historical textual data is now available for research on numerous topics including historical linguistics and conceptual history. We summarise previous research which has shown that it is necessary to map historical spelling variants to modern equivalents in order to successfully apply natural language processing and corpus linguistics methods. Manual and semi-automatic methods have been devised to support this normalisation and standardisation process. We argue that it is important to develop a linguistically meaningful rationale to achieve good results from this process. In order to do so, we propose a number of guidelines for normalising corpora and show how these guidelines have been applied in the Corpus of English Dialogues.

1 Introduction

The development of (semi-)automatic tools such as the VARiant Detector, henceforth VARD (Baron and Rayson 2008, 2009), has afforded compilers of historical corpora the opportunity to normalise variant spellings relatively quickly, following a dedicated period of manual training using corpus samples (see, e.g., Lehto *et al*. 2010). In the case of VARD, for example, this period of manual training involves the user:

(i) reading a given text, via the VARD graphical user interface,
(ii) distinguishing variants within the text – via the tool's automatically recommended set of candidate variants – or personally – by highlighting variant forms manually,

(iii) choosing the most appropriate normalised form for each variant found; where relevant, being guided by VARD's recommended list of candidate replacements ranked by an f-score calculation (derived from, e.g., known variants, letter replacement rules, edit distance measures and/or phonetic matching algorithms),
(iv) matching the variant with the normalised form – but in such a way that the original spelling is retained in an XML tag (Baron and Rayson 2008).

The argument for normalisation is twofold. First, that it helps to improve the accuracy of automated computational linguistic (natural language processing) techniques such as part-of-speech tagging and second, that it improves the stability and robustness of corpus linguistic methods such as keyword analysis, thereby allowing existing software tools of both types to be used unmodified (Archer *et al*. 2003; Rayson *et al*. 2007a, b; Baron *et al*. 2009; Hiltunen and Tyrkkö 2013). It goes without saying that such normalisation needs to be handled sensitively: so that, for example, we can maintain – within the text – the original spelling of those forms which convey important morphosyntactic or orthographic information (as opposed to retaining these original spellings as part of the XML tag – see (iv), above). Hence the inclusion of an IGNORE VARIANT facility within VARD. In this paper, we describe the decisions we have made with respect to the *Corpus of English Dialogues 1560–1760*, when determining which features require normalisation and which should be left as they were originally (and why). In particular, we discuss our treatment of names; the genitive construction; auxiliaries and verbs; (open-hyphenated-closed) compounds; abbreviations; graphemes such as the tilde; terms which are now archaic, obsolete or rare; foreign terms; dialect terms; and personal pronouns (see Sections 4.1–4.3).

Published in 2006, the compilation of the *Corpus of English Dialogues* (CED) represents a cross-University collaboration between Uppsala and Lancaster, made possible thanks to grants from the Swedish Research Council, the Arts and Humanities Board and the British Academy. The corpus totals 1,157,720 words, and covers a 200-year period (1560–1760), divided into five 40-year sub-periods: of these 870,240 words have been coded for direct speech. Each 40-year sub-period contains speech-related texts representative of five genres – the courtroom, witness proceedings, comedy dramas, prose fiction and handbooks; the first four sub-periods also contain a group of texts subsumed under a 'miscellaneous' category. The CED thus makes possible speech-related studies using historical pragmatic-, historical sociopragmatic- and variationist frameworks.

This research, although focussed on the CED in this paper, has an additional, wider aim: determining the feasibility of developing normalisation guidelines that are generalisable to other historical corpora such as ARCHER (*A Representative Corpus of Historical English Registers*) and EEBO (*Early English Books Online*). As part of the Semantic Annotation and Mark-up for Enhancing Lexical Searches (SAMUELS) project, for example, Archer, Baron and Rayson are training VARD on 25-year sub-corpora, taken from EEBO, in order to obtain specialised models for different time periods. Funded by the Arts and Humanities Research Council (AHRC) in conjunction with the Economic and Social Research Council (ESRC), the wider context of the SAMUELS project is to build a Historical Thesaurus Semantic Tagger, thereby giving users a system for automatically annotating words in texts with their precise meanings (where necessary, disambiguating between possible meanings of the same word).

This desire to determine the feasibility of developing normalising guidelines that are generalisable across a range of historical corpora motivates our comparison of the normalisation decisions we made in respect to the *Corpus of English Dialogues* (CED), in Sections 4.1–4.3, with those made by Lehto *et al*. (2010) in respect to the corpus of *Early Modern English Medical Texts* (EMEMT); for the normalisation of the Corpus of Early English Correspondence (CEEC), see Palander-Collin and Hakala (2011). We begin, however, with a brief summary of the extent of spelling variation in the Early Modern English (EModE) period – as evidenced in available EModE corpora (see Section 2, following), before going on to explain the VARD tool in more detail, and the motivations for its development (Section 3).

2 The extent of spelling variation in EModE corpora

Prior to the development of tools such as VARD, researchers tended to adopt a qualitative approach in order to study spelling variation (but see Schneider 2002 who adopts a corpus-based approach to study a restricted time-period). Studies worthy of mention here – because of their focus on 1500–present – include Elphinston (1765, 1790), Walker (1791), Wyld (1923, 1927, 1936), Kökeritz (1953), Dobson (1955, 1957), Scragg (1974), Cercignani (1981), Blake (1989, 2002), Jones (1989), Görlach (1991), Lass (1999), Rissanen (1999), Salmon (1999), Beal (2002, 2006), Sebba (2007), Sairio (2009) and Evans (2012). Research studies interested in standardisation will also tend to pick up on spelling variation. Often, this is to argue how,

> […] the English spelling system that emerged from the seventeenth century is not a collection of random choices from the ungoverned mass of alternatives that were available at the beginning of the century but rather a highly ordered system taking into account phonology, morphology and etymology and providing rules for spelling the new words that were flooding the English lexicon. Printed texts from the period demonstrate clearly that, during the middle half of the seventeenth century, English spelling evolved from near anarchy to almost complete predictability (Brengelman 1980: 334).

Other researchers – most notably, Osselton (1963, 1984) and Tieken-Boon van Ostade (1998) – have suggested that the eighteenth century is worthy of special attention from those interested in the history of English spelling, not least because of the (at times) considerable differences between private spelling habits and the spelling of printed texts – such that we might talk in terms of "a public spelling system and a private one" (Tieken-Boon van Ostade 1998: 457; see also Evans 2012). Van Ostade further suggests that, by "the end of the eighteenth century, the printers' spelling had evidently established itself as the only correct way to spell" and also that "the ability to spell correctly had acquired social significance/ by that time" (van Ostade 1998: 466–467). Researchers such as Carney (1994: 467), in contrast, have argued that (spelling) "standardization was only indirectly the work of printers", not least because English spelling "was too well-designed to be a simple settling down of printing house practices".

Our ability to document this apparent decline in variant spellings, across the EModE period, is much easier, thanks to VARD, in conjunction with DICER, a web-based tool for exploring spelling patterns. Specifically, DICER analyses the XML tagged output from VARD and extracts the letter edit rules which transform variants to their corresponding normalised form, e.g. "remove final *e*", "substitute initial *v* for *u*", and "substitute *y* for *i*", and produces frequencies linked to metadata.[1] These edit rules can then be used to determine the most frequent – as well as the more unusual – patterns relating to spelling variation. With VARD, we can also identify any decline in variant spellings across time and, importantly, across text-types or genres (due to the growing availability of EModE corpora; but see Section 5). For example, Figure 1, taken from Baron *et al*. (2009), shows the extent of non-standard spellings in six corpora representative of the EModE period – ARCHER, EEBO, Innsbruck, Lampeter, EMEMT and Shakespeare. In fact, variation is revealed to be characteristic of all six corpora throughout this period, though the extent of variation becomes less and less as we reach the end of the eighteenth century.

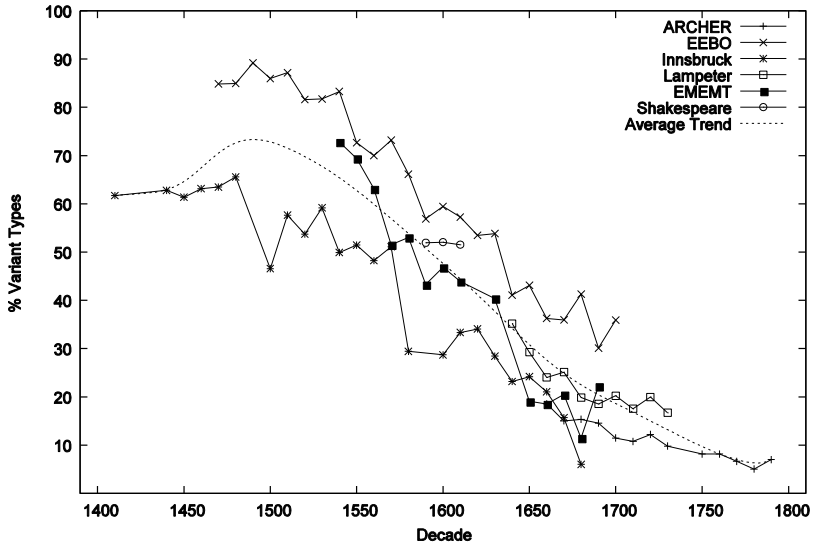

Figure 1: Percentage of variant types in six EModE corpora

Such spelling variation is known to directly affect corpus and computational linguistics methodologies. By way of illustration, because word frequency lists show multiple variant spellings instead of one form, the *concordancing* process is of little help unless – at the point of undertaking an investigation relating to *would*, for example – the user knows (to check for occurrences of) its related variant forms, for example, *wold*e, *woolde*, *wuld*, *wulde*, *wud*, *wald*, *vvould*, *vvold*, etc.

The key words procedure is also affected, as revealed by Figure 2 (taken from Baron *et al*. 2009) which shows the extent of difference in keyword ranking with and without normalisation (where 1 equates to the same list, and anything below 1 demonstrates variation across keywords; the dotted line shows the smoothed trend from the actual points on the line).[2] Others have shown similar effects on key word clusters (Palander-Collin and Hakala 2011).

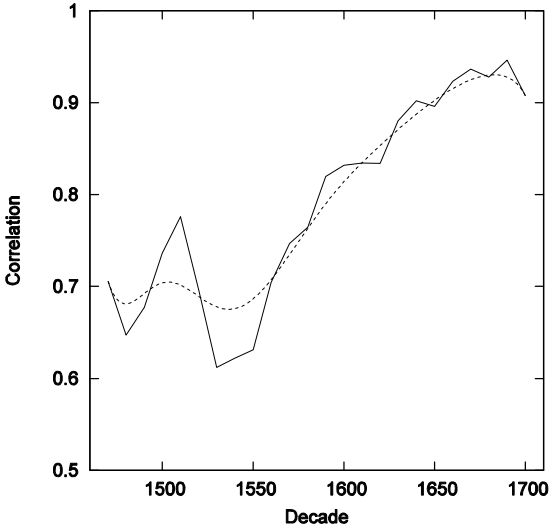

Figure 2: Correlation of normalized and non-normalised keyword lists for EModE corpora

3 Development of VARD

The original motivation behind VARD was the creation of a pre-processing step which would improve the accuracy of automatic part-of-speech tagging and semantic annotation processes applied to EModE datasets (Archer *et al*. 2003; Rayson *et al*. 2007a/b). As part of this development, Archer and Rayson have worked with others to assess the usefulness of existing modern spell-checking techniques (Rayson *et al*. 2005); and, when joined by Baron, have adopted hybrid methods (Baron and Rayson 2008) to, first, detect historical spelling variants and, second, suggest modern equivalents with which to link/replace them. VARD uses a large modern dictionary or word list derived from the *British National Corpus* (BNC) and other sources as a reference list against which to compare each word in a historical text. Any words from the corpus that do not appear in the word list are flagged as potential historical variants. Four methods are then used to suggest candidate modern matches for each variant:

1. A known variants list consisting of historical and modern pairs, which has been manually created and extended by a user selecting or inserting a matching modern form into the interactive version of VARD.
2. Letter replacement rules (such as *u* to *v* and *ie* to *y*), derived from existing literature or corpora that have been manually VARDed, are used to transform the historical variant to the modern form.
3. A phonetic matching algorithm (a variant of SoundEx), used to uncover similar historical and modern forms.
4. The Levenstein edit distance metric, used to measure the number of character insertions, deletions and substitutions required to transform the word from the historical variant to the modern equivalent. This along with precision and recall scores and frequency in the BNC is used to rank the potential modern equivalents.

Ongoing work in the SAMUELS project is also deriving improved metrics and rules from the variants, headwords and dates in the *Oxford English Dictionary.*

4 VARDing issues relating to the Corpus of English Dialogues (CED)

As previously highlighted, the latest version of the tool, VARD2[3], is designed to assist researchers in standardising spelling variation in historical corpora both manually and automatically, thereby enabling the use of standard corpus and computational linguistics tools without any modification. Specifically, the tool draws on methods from modern spellchecking to find spelling variants and offer/select appropriate modern equivalents, but in such a way that the original spelling is retained in the text with an XML tag surrounding the replacement. Hence:

<normalised orig="charitie">charity</normalised>.

The tool has already been used to normalise EMEMT (Lehto *et al.* 2010) and the *Corpus of Early English Correspondence* (CEEC) (Palander-Collin and Hakala 2011), allowing us to compare our decisions for (not) normalising particular variants with their motivations, where relevant (see Sections 4.1–4.3).

The VARD2 developers maintain that optimum results are achieved when users first undertake a period of manual training as outlined in the introduction to this paper. This means reading a given text from a training set, via the VARD interface; highlighting variant forms manually and/or by allowing the tool to highlight them; then choosing whether to (i) leave the highlighted variant form

as is, (ii) keep the form but normalise the spelling or (iii) modernise the form. When taking options (ii) or (iii), the user should select the most appropriate form matching the variant, being guided (where relevant) by VARD's recommended list of candidate replacements. These are ranked by an f-score calculation which is based on a combination of methods: known-variants list, letter replacement rules, edit distance measures and/or phonetic matching algorithms (Baron and Rayson 2008). We have adopted this approach when normalising CED. Our particular training set consists of twenty-five 1,200-word chunks (taken from twenty-five of the CED's 177 files), and totals 30,213 words (having preserved full sentences that go beyond a chunk boundary). This training set is made up of five files per each of the five sub-periods to ensure it is representative of trials, depositions, comedy dramas, prose fiction and handbooks. The EMEMT training data set is comparable, comprising 36,000 words (preceded by an initial training set of 24,000 words). The two-million-word EMEMT corpus contains 450 texts/text samples from the period 1500–1700 (see Taavitsainen and Pahta 2010, and http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/EMEMTindex.html).

In the remaining sections, we discuss the decisions we have made in respect to (i), (ii) and (iii), namely:

(i) Leaving a word form as it was.
(ii) Keeping the form but normalising its spelling into one form across the spelling variants.
(iii) Modernising the form.

Where relevant, we also comment on the similarities/differences between our principles adopted for normalising spelling variation in the CED and those adopted for EMEMT. This will help to make clear when/where our decisions followed the principles applied to these corpora and when/where we opted for different solutions.

4.1 Variants treated with caution

Similar to the practices adopted for the EMEMT corpus, we decided to treat with caution most names (e.g. *Darbye, North Baiely*), archaisms or obsolete terms (e.g. *cozen/ed*, *oft*, *morrow*) and leave them as they were (i.e., adopted method (i); although we did opt for method (ii) in some cases, that is, we normalised spellings such as *ofte* to *oft*. Correspondingly, foreign and dialectal terms (*birlady* > *byr'lady*), personal pronouns (*thyne* > *thine*), and archaic or obsolete forms were normalised – but to one historical variant spelling only (cf. being

replaced by their modern alternatives). Latinate, foreign and dialectal forms were standardised (but not written out as in e.g. *by our lady*, cf. above). The forms of the second-person singular pronouns were kept and only standardised in spelling, as replacing them by modern forms would have changed their possible connotations in use (cf. Lehto *et al*. 2010: 286).

4.2 Variants that benefit from modernisation

Among the forms that we believe would benefit from modernising, we count genitive forms, auxiliaries, verbs, compounds, contractions, tilde and other special characters used for abbreviations as described, in turn, in each of the following subsections.

4.2.1 Genitive (and other uses of the apostrophe)

Having suggested that genitive forms are conducive to method (iii), we should point out that genitive forms can nonetheless prove to be a challenge in some cases; especially given our preference for singular forms to be distinguished from plural forms, where context allows such a distinction to be made. Such preferences suggest that a fully-automated approach to spelling normalisation may not be advisable, or, at the very least, that a user considers what morpho-syntactic information might be lost, thereby making certain linguistic studies inadvisable using automatically-normalised datasets.

Specific examples (from a 1599 play by George Chapman) required rather radical modernization:

(1) then may you well say, seeing my race is so profitably increased, that good fat oxe, and that same large eard asse are *my sonne sonnes*, that caulfe with a white face is his faire daughter, (D1CCHAPM)

(2) [$ (^Lab.^) $] Talke not to me of creame, for such vaine meate I do despise as food, my stomack dies drowned in the cream boules of *my mistres eyes*. (D1CCHAPM)

My sonne sonnes was modernized manually to *my son's son*; *my mistres eyes* to *my mistress's eyes*. Once identified, such forms can be added to the VARD's known variant list. Other examples of modernisations applied to the rendering of the apostrophe, which can be added to the VARD's known variant list, include *giue's* to *give us* and *Ille* to *I'll*.

4.2.2 Auxiliaries and verbs

The categories where normalisation looked like a natural procedure to adopt included verb endings, among them the past tense or past participle *-t* and *'d*,

third-person singular *-th* and second-person singular *-st*, e.g. the forms given in examples (3)–(5):

(3) "[...] at this the King **<normalised orig="laught" auto="false">laughed</normalised>**[...]" (D2FARMIN)

(4) "Then she desired the following Witnesses might be **<normalised orig="call'd" auto="false">called</normalised>** in her Defence." (D5WBLAND)

(5) "An unlikely matter; but thus you see the Duke **<normalised orig="confesseth" auto="false">confesses</normalised>** the receipt of the Letter[...]" (D1TNORFO)

These were normalised into modern forms; thus, forms such as *shew*, *shews* and *shewed* were given as *show*, *shows* and *showed*, *wouldst, wouldest, would'st* were rendered as *would*, and *didst* and *dost* as *did* and *do*. In some respects, the decisions taken for the CED appear more radical than those adopted for the EMEMT corpus, which has kept e.g. *dost* and *doest* (but replaced e.g. *didst* and *diddst* by *did*). Similarly, *doth* and *hath* were kept in the EMEMT corpus as they were used for both the singular and plural in the corpus texts; these forms have been normalised in the CED by using modern forms, although some manual screening still remains to be done on the basis of concordances to tell apart the singular and plural uses.

4.2.3 Compounds

Compound forms can present problems, as the use of the space between the elements may have fluctuated across centuries. The main principle adopted for normalising compound spellings in the CED was to split or divide the words as in Present-day English; however, problematic cases were left as they were in the original CED. Thus pronouns such as *my self* and *your self* were rendered as *myself* and *yourself*. The line taken for the EMEMT corpus, in contrast, was not to interfere in word divisions at this level. Other instances normalised in the CED included *any way* as *anyway, to morrow* as *tomorrow, shalbe* as *shall be* and *an other* as *another*. For corpus examples, see (6)–(8) below:

(6) "Pray don't trouble **<join original="your self">yourself</join>** on my Account." (D5HGBEIL)

(7) "And, if you please, **<join original="to morrow">tomorrow<join>** we shall begin." (D4HEMIEG)

(8) "It **<normalised orig="shalbe" auto="false">shall be</normalised>** then for **<join original="an other">another</join>** <normalised orig="tyme" auto="false">time</normalised>." (D1HEBELL)

As joining words separated by a space is something that VARD2 does not know how to deal with automatically (the program does not take the context into account), these instances will need to be dealt with manually (i.e. using the join tag within VARD's interactive mode, as indicated above).

4.2.4 Contractions

As for contractions and other abbreviated forms, we normalised the items where the corresponding full forms of the expressions were known in Present-day English. Among the examples illustrating this principle are the following:

'em	>	*them*
for'it	>	*for it*
igad	>	*i'gad*
on't	>	*on it*
sblood	>	*s'blood*
sha'n't	>	*shan't*
tho	>	*though*
tis or 'tis	>	*it's*
twas, t'was	>	*it was*
twill, t'will	>	*it'll*
qd	>	*quod*
weel(e)	>	*we'll*
wy	>	*with you*
y'are	>	*you're*
yfaith, yfayth, ifaith	>	*I'faith*

On this point the practices adopted for individual items may differ between the CED and the EMEMT.

4.2.5 Tilde

The tilde was commonly used in abbreviations in Early Modern English to mark the nasal consonants *n* and *m*. We normalised these instances into corresponding Present-day forms, as in examples (9) and (10).

(9) Let <normalised orig="vs" auto="false">us</normalised> begin **<normalised orig="the~" auto="false">then</normalised>**. (D1HEBELL)

(10) But you dealt all to **<normalised orig="the~" auto="false">them</normalised>**. (D1HEBELL)

In fact, our training data only contained four instances of the tilde standing for *m*, all of them in the pronoun *them*; in the other instances with the tilde, the special character stood for the letter *n* (e.g. *the~, vpo~, husba~ds, we~t, ma~, informatio~, dispositio~*). In the EMEMT corpus, tildes were also replaced by e.g. *n* or *m*, but occurrences of non-replaced tildes still occur in the texts (Lehto *et al*. 2010: 286).

4.3 Context-based decisions

When screening instances of tilde, checking the context proved crucial for making decisions about how to normalise the forms. Other common uses requiring contextual scrutiny in Early Modern English include *bee/be, doe/do, the/thee, then/than*, *to/too*, and *yt/y=t=/that* (*y=t=* in a corpus text standing for y^t in the original). With the verb forms *be/bee*, the EMEMT team first automatically normalised all instances of *bee/be* into *be* and reversed the few that stood for the noun *bee* by screening concordances. All our instances of *bee* were missed normalisations of the verb *be*; the same held for the instances of *doe*. For corpus examples, see (11)–(15):

(11) the more it is to *bee* feared?
> the more it is to **<normalised orig="bee" auto="false">be</normalised>** feared? (D2FARMIN)

(12) What to *doe*?
> What to <normalised orig="doe" auto="false">do</normalised>? (D1HEBELL)

(13) and make *the* spend all thie meanes.
> and make **<normalised orig="the" auto="false">thee</normalised>** spend <normalised orig="thie" auto="false">thy</normalised> whole estate" (D2WDIOCE)

(14) Excuse me, Sir, I understand it more *then* I do high German.
> Excuse me, Sir, I understand it more **<normalised orig="then" auto="false">than</normalised>** I do high German." (D3HFFEST)

(15) in good faith you are *too* blame
> in good faith you are **<normalised orig="too" auto="false">to</normalised>** blame [...]" (D1CHAPM)

As for *y=t=/yt* standing for *that*, we decided to normalise the only instance of *y=t=* occurring in our training set:

(16) hir husbande said diuers times *y=t=* he would cut it of,
> <normalised orig="hir" auto="false">her</normalised> <normalised orig="husbande" auto="false">husband</normalised> said <normalised orig="diuers" auto="false">divers</normalised> times **<normalised orig="y=t=" auto="false">that</normalised>** he would cut it <normalised orig="of" auto="false">off</normalised>,

However, superscript forms were left unchanged in the normalised version of the EMEMT.

5 Concluding comments, and studies made possible

In this paper, we have described our work towards defining a set of guidelines for the normalisation of historical EModE corpora. A larger motivation for offering a set of guidelines for the normalisation of historical corpora, however, is that we believe that tools like VARD2 can begin to in/validate the various established theories in respect to the motivations for the decline in spelling variation (see, e.g., Section 2), by affording – for the first time – a quantitative approach to the study of spelling variation over time. Questions which we (and hopefully others) can pose include: whether/to what extent the introduction of printing into England in the late 15th century promoted new orthographic practices – and, if so, which ones specifically? How long did typesetters continue to allow for spelling variation after the advent of printing – and were there regional (or idiolectal) differences in terms of the choices they made (as there were in the Middle English period)? What society-shaping events (if any) affected spelling variation from 1500 onwards? The Civil War (1642–9) perhaps? What is the relationship between linguistic change and spelling standardisation over time? And what part (if any) was played by the Great Vowel Shift (circa 1350–1500)? Our reviews, to date for example, suggest that researchers have not tended to link the Great Vowel Shift specifically to spelling standardisation – beyond, perhaps, Stenbrenden (2010)[4] – but is this an oversight? This demands, in turn, that we give further consideration to the relationship between phonology and orthography: what should we make, for example, of Smith's (1996: 23) claim that "…informal writings from the [Early Modern English] period…sporadically reflect[ed] contemporary speech-habits" but "in ways which [tend to be] disguised in contemporary printed books". For Smith, such phenomena "deserve more attention than they are given" by students of EModE speech, which means

looking beyond "the works of orthoepistical writers and the early English grammarians and lexicographers" (Smith 1996: 23). Smith is alluding to a much more popular thesis related to spelling, of course: that of the role played by dictionaries as instruments of standardisation, which, in turn, precipitated a period of linguistic prescriptivism, resulting in a steady decline in variant spellings in printed texts from the mid-seventeenth century onwards (Brengelman 1980: 334). In which case, it is helpful to explore the role played by specific spelling reformers such as James Elphinston (1721–1809); but using a combination of quantitative as well as qualitative and/or (semi) automatic as well as manual approaches to study spelling variation, such that we detect – and, hence, better understand – their contemporary influence as well as any lasting influence on (the 'fixing' of) English spelling over time. Relatedly, more extensive research, using computational tools like VARD2, would enable researchers to draw on large datasets to (in)validate the theory that spelling irregularity did decline rapidly after the mid-1700s, as Brengelman (1980) suggests. Researchers might search, in turn, for evidence of the divergence between the public and private standard in relation to spelling – as a means of determining whether such "epistolary spelling" (Tieken-Boon van Ostade 1998: 467) survived until the late eighteenth century, as has been argued previously. But what should we make of warnings such as those given by Rissanen (1998), Nevalainen and Raumolin-Brunberg (2003) and Evans (2012)? Rissanen points out, for example, that:

> …historical corpora should never be used as an excuse for overlooking the study of primary texts. Sophisticated computer technology and multifactor analysis are useful only when combined with a profound knowledge of the language form and period under scrutiny (Rissanen 1998: 390).

Evans, quoting Nevalainen and Raumolin-Brunberg, addresses spelling studies specifically, and suggests that there is a problem beyond the method of retrieval – that of finding suitable data for analysis:

> Nevalainen and Raumolin-Brunberg (2003: 44) suggest that the mixed origin of the transcriptions (apograph manuscripts, published collections) make CEEC unsuitable for the scrupulous study of spelling. The uncertain authenticity of copied texts, which may contain such minor differences as the omission or insertion of a final <e>, means that the social background of the claimed author and the spelling features cannot be reliably correlated. Furthermore, different editorial practices implemented in the print versions of the texts lead to uncertainties over

> possible silent corrections for clarity, for instance, or discrepancies in the editors' reading of the manuscript (see Smith and Kay 2011 for an exploration of these issues in relation to Older Scots poetry) (Evans 2012).

We agree with such statements. But, as with all historical studies, there are occasions when we have to work with what we have – whilst using that data sensitively, in ways that demonstrate both (i) an awareness of the period under scrutiny, and (ii) a transparency in respect to the strengths and weaknesses of the materials and methods used. Rather than invalidating the creation of VARD, however, we would argue that it motivates any ongoing development of such tools (as a means of further refining their efficacy in respect to interrogating and retrieving data). A need to know one's data – and how to treat it to enable various studies (be they historical sociolinguist, pragmatic, lexicographical, variationist, etc.) – also validates the need for guidelines for normalising EModE corpora. Hence, our work towards such guidelines here. For example, we would argue that the guidelines are necessarily a compromise between full modernisation of all word forms, normalisation to a 'standard' EModE variant and leaving certain variants as per the original spelling. In creating the guidelines, we have taken account of sensitivities related to applications in historical linguistics, conceptual history, and corpus and computational linguistics. Our work on the CED training set, and the comparisons we made with the EMEMT corpus, confirmed that normalisation guidelines can vary and that they are often subject to the amount of data requiring manual screening and the amount of resources available to carry out the manual work. It is also important to combine automatic processing and manual screening in the normalisation process. In the near future, we will apply VARD2 to the EEBO corpus (minus the 25-year-period subcorpus currently being used by Archer, Baron and Rayson for training purposes) to enable much more refined time-sensitive normalisations: this will allow us to determine whether our guidelines are robust enough when applied on a much larger scale to billion-word sized corpora.

Acknowledgements

We gratefully acknowledge the support of the SAMUELS project funded by the AHRC (grant reference AH/L010062/1), see http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/. We are also indebted to The Royal Society of Arts and Sciences of Uppsala for a travel grant enabling us to

arrange planning and training sessions. Thanks, in addition, are due to Terry Walker and Gerold Schneider for acting as VARDers on the CED in June 2013.

Notes

1. For more details in respect to DICER, see http://corpora.lancs.ac.uk/dicer/.
2. However, as the correlation coefficient reveals, keywords are less affected as we approach 1700 (thanks to the reduction in variation reduces).
3. VARD2 is freely available for academic use: for details, see http://ucrel.lancs.ac.uk/vard/.
4. Stenbrenden (2010) investigates the phonological changes of long monophthongs through the development of spelling variation, but focuses on the Middle English period (c. 1100–1500).

References

Archer, Dawn, Anthony M. McEnery, Paul Rayson and Andrew Hardie. 2003. Developing an automated semantic analysis system for Early Modern English. In D. Archer, P. Rayson, A. Wilson and A. M. McEnery (eds.). *Proceedings of the Corpus Linguistics Conference 2003*, 22–31. Lancaster: University of Lancaster.

Baron, Alistair and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, 22 May 2008. See http://eprints.lancs.ac.uk/41666/1/BaronRaysonAston2008.pdf

Baron, Alistair and Paul Rayson. 2009. Automatic standardization of texts containing spelling variation, how much training data do you need? In M. Mahlberg, V. González-Díaz and C. Smith (eds.). *Proceedings of the Corpus Linguistics Conference, CL2009,* University of Liverpool, UK, 20–23 July 2009.
See http://ucrel.lancs.ac.uk/publications/cl2009/314_FullPaper.pdf

Baron, Alistair, Paul Rayson and Dawn Archer. 2009. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies* 20 (1): 41–67.

Beal, Joan C. 2002. *English pronunciation in the Eighteenth Century: Thomas Spence's Grand Repository of the English Language*. Oxford: Oxford University Press.

Beal, Joan C. 2006. *Language and region.* London and New York: Taylor & Francis.

Blake, Norman. 1989. *The language of Shakespeare.* Houndmills, Basingstoke, Hampshire and London: Macmillan.

Blake, Norman. 2002. *A grammar of Shakespeare's language.* Houndmills, Basingstoke, Hampshire and London: Palgrave.

Brengelman, Fred. H. 1980. Orthoepists, printers and the rationalisation of English spelling. *Journal of English and Germanic Philology* 79: 332–354.

Carney, Edward. 1994. *A survey of English spelling*. London and New York: Routledge.

Cercignani, Fausto. 1981. *Shakespeare's works and Elizabethan pronunciation.* Oxford: Clarendon Press.

A Corpus of English Dialogues 1560–1760. 2006. Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University), with the assistance of Dawn Archer and Terry Walker.

Dobson, Eric J. 1955. Early Modern Standard English. *Transactions of the Philological Society,* 25–40.

Dobson, Eric J. 1957. *English pronunciation 1500–1700.* Oxford: Clarendon Press.

Elphinston, James. 1765. *The principles of the English language digested: or, English grammar reduced to analogy…* 2 vols. London. A i.261.

Elphinston, James. 1790. *Inglish orthoggraphy epitomized …* London. EL 288. A vi.544.

Evans, Mel. 2012. A sociolinguistics of early modern spelling: An account of Queen Elizabeth I's correspondence. In J. Tyrkkö, M. Kilpiö, T. Nevalainen and M. Rissanen (eds.). *Outposts of historical corpus linguistics: From the Helsinki Corpus to a proliferation of resources* (Studies in Variation, Contacts and Change in English 10 [online.]). Available at: <http://www.helsinki.fi/varieng/series/volumes/10/evans/#taavitsainen_2000>
[Last accessed 09/12/2014].

Görlach, Manfred. 1991. *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Hiltunen, Turo and Jukka Tyrkkö. 2013. Tagging *Early Modern English Medical Texts*. Corpus Analysis with Noise in the Signal (CANS) 2013 workshop. Lancaster University. See http://ucrel.lancs.ac.uk/cans2013/

Jones, Charles. 1989. *A history of English phonology.* London: Longman.

Kökeritz, Helge. 1953. *Shakespeare's pronunciation.* New Haven: Yale University Press.

Lass, Roger. 1999. Introduction. In R. Lass (ed.), *The Cambridge history of the English language: Volume III. 1476–1776,* 1–12. Cambridge: Cambridge University Press.

Lehto, Anu, Alistair Baron, Maura Ratia and Paul Rayson. 2010. Improving the precision of corpus methods: The standardized version of *Early Modern English Medical Texts*. In I. Taavitsainen and P. Pahta (eds.). *Early Modern English Medical Texts: Corpus description and studies*, 279–290. Amsterdam: John Benjamins.

Nevalainen, Terttu and Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England*. (Longman Linguistics Library). London: Longman Pearson.

Osselton, Noel E. 1963. Formal and informal spelling in the 18th century. *Errour*, *honor* and related words. *English Studies* 44: 267–275.

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

Palander-Collin, Minna and Mikko Hakala. 2011. Standardizing the *Corpus of Early English Correspondence* (CEEC). A poster given at the 32nd ICAME conference, 1–5 June, 2011. See http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/standardized.html;
for an enlarged version, see http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/Standardization%20poster%20v2.pdf.

Rayson, Paul, Dawn Archer and Nick Smith. 2005. VARD versus WORD: A comparison of the UCREL variant detector and modern spellcheckers on English historical corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham University, July 14–17, 2005.

Rayson, Paul, Dawn Archer, Alistair Baron and Nicholas Smith. 2007a. Tagging historical corpora – the problem of spelling variation. In *Proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491*, International Conference and Research Center for Computer Science, Schloss Dagstuhl, Wadern, Germany, 3rd–8th December 2006. ISSN 1862-4405. http://www.comp.lancs.ac.uk/~paul/publications/rabs_extAbs_dagstuhl06.pdf

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nicholas Smith. 2007b. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of the Corpus*

Linguistics Conference 2007. Birmingham: University of Birmingham. http://comp.eprints.lancs.ac.uk/1528/1/192_Paper.pdf.

Rissanen, Matti. 1998. Towards an integrated view of the development of English: Notes on causal linking. In J. Fisiak and M. Krygier (eds.). *Advances in English historical linguistics*, 389–406. Berlin: Mouton de Gruyter.

Rissanen, Matti. 1999. Syntax. In R. Lass (ed.). *The Cambridge history of the English language: Volume III. 1476–1776*, 187–331. Cambridge: Cambridge University Press.

Sairio, Anni. 2009. *Language and letters of the Bluestocking Network: Sociolinguistic issues in eighteenth-century epistolary English* (Mémoires de la Société Néophilologique de Helsinki 75). Helsinki: Société Néophilologique.

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

Schneider, Peter. 2002. Computer assisted spelling normalization of 18th century English. In P. Peters, P. Collins and A. Smith (eds.). *New frontiers of corpus research: Papers from the 21st International Conference on English Language Research on Computerized Corpora, Sydney, 2000*, 199–211. Amsterdam: Rodopi.

Scragg, Donald C. 1974. *English spelling*. Manchester: Manchester University Press.

Sebba, Mark. 2007. *Spelling and society: The culture and politics of orthography around the world.* Cambridge: Cambridge University Press.

Smith, Jeremy. 1996. *A historical study of English: Form, function and change*. London: Routledge.

Stenbrenden, Gertrud. 2010. The chronology and regional spread of long-vowel changes in English, c. 1150–1500. PhD dissertation, University of Oslo.

Taavitsainen, Irma and Päivi Pahta (eds.). 2010. *Early Modern English Medical Texts. Corpus description and studies*. Amsterdam/Philadelphia: John Benjamins.

Tieken-Boon van Ostade, Ingrid. 1998. Standardization of English spelling: The eighteenth-century printers' contribution. In J. Fisiak and M. Krygier (eds.). *Advances in English historical linguistics*, 457–470. Berlin: Mouton de Gruyter.

Walker, John 1791. *A critical pronouncing dictionary and expositor of the English language.* London.

Wyld, Henry C. 1923. *Studies in English rhymes from Surrey to Pope.* London: Murray.

Wyld, Henry C. 1927. *A short history of English*. 3rd edition. London: Murray.

Wyld, Henry C. 1936. *A history of modern colloquial English*. 3rd edition. Oxford: Basil Blackwell.