sciendo

ICAME Journal, Volume 45, 2021, DOI: 10.2478/icame-2021-0002

Supporting the corpus-based study of Shakespeare's language: Enhancing a corpus of the First Folio

Jonathan Culpeper, Andrew Hardie, Jane Demmen, Jennifer Hughes and Matt Timperley, Lancaster University

Abstract

This article explores challenges in the corpus linguistic analysis of Shakespeare's language, and Early Modern English more generally, with particular focus on elaborating possible solutions and the benefits they bring. An account of work that took place within the Encyclopedia of Shakespeare's Language *Project (2016–2019) is given, which discusses the development of the project's data resources, specifically, the* Enhanced Shakespearean Corpus. *Topics covered include the composition of the corpus and its subcomponents; the structure of the XML markup; the design of the extensive character metadata; and the word-level corpus annotation, including spelling regularisation, part-of-speech tagging, lemmatisation and semantic tagging. The challenges that arise from each of these undertakings are not exclusive to a corpus-based treatment of Shakespeare's plays but it is in the context of Shakespeare's language that they are so severe as to seem almost insurmountable. The solutions developed for the* Enhanced Shakespearean Corpus *– often combining automated manipulation with manual interventions, and always principled – offer a way through.*

1 Introduction

The works of William Shakespeare have received plenty of attention from literary critics, but much less attention from linguists. Despite advances in digital methods, no study of Shakespeare's language has deployed the corpus-based methods used today by, for example, lexicographers and grammarians in a comprehensive way. Of course, the rapid rise of digital humanities has produced related endeavours. For example, computational techniques in studies of authorship attribution are now quite familiar within Shakespearean scholarship and have contributed greatly to it. For example, Craig and Greatley-Hirsch (2017) use statistical techniques (including principal components analysis) to illumi-

nate issues such as the role of prose and verse, characterisation, the use of stage props, dramatic genre and authorial style. However, the computational and quantitative techniques utilised in authorship attribution research are quite distinct from those which underpin corpus linguistics, including but not limited to the basic conceptualisation of language which they assume. Thus, for instance, while the patterns that Craig and Greatley-Hirsch identify are sensitively interpreted, they have been calculated from counts of a purposefully simplistic 'unit' of language, namely the decontextualised word, across texts or sub-text segments each modelled as without structure, that is, as a 'bag of words'. But words *do* form structures, and those structures *do* have meaning – points which are among the fundamental assumptions of corpus linguistic methods, as evident in, for example, the notion and operationalisation(s) of collocation as a mainstay of both theoretical and applied language research (McEnery and Hardie 2012: 122ff). Addressing collocations means identifying (quantitatively) unusual patterns of word co-occurrence; but, crucially, it also requires the *interpretation* of these co-occurrence patterns as components within the language, in terms either of system or of discourse – be that as chunks of language that form the lexicogrammar, or as conventionalised units of semantic-pragmatic expression (see, for example, Gledhill 2000). Other utilities in the corpus linguistic toolbox also unite computational and/or statistical information with linguistically informed theoretical apparatuses. A plethora of research, such as authorship attribution, applying non-corpus-based digital methods to the works of Shakespeare still, then, leaves much productive work to be done within the specific framework of the corpus methodology.

The AHRC-funded *Encyclopedia of Shakespeare's Language* Project (2016–2019) set out to fill this gap, and thus to bring scholarship on Shakespeare's language fully into the twenty-first century. As its title indicates, this project's major output is the five-volume *Encyclopedia*, to be published by Bloomsbury under the Arden umbrella, which incorporates a dictionary in its first two volumes, with further volumes presenting linguistic profiles of characters and plays, networks of characters, and themes. All flow from the analysis of a corpus specifically created to assist the study of Shakespeare's language, the *Enhanced Shakespearean Corpus*. This consists of three primary and two secondary components. The primary components are:

Enhanced Shakespearean Corpus: First Folio Plus (hereafter, *ESC:Folio*)
Enhanced Shakespearean Corpus: Comparative Plays (hereafter, *ESC:Comp*)
Enhanced Shakespearean Corpus: EEBO-TCP Segment (hereafter, *ESC:EEBO*)

The secondary components contain further texts by Shakespeare: his published non-theatrical poetry, and earlier quarto editions of certain plays which precede the First Folio edition on which *ESC:Folio* is largely based:

Enhanced Shakespearean Corpus: Quartos (hereafter, *ESC:Quartos*)
Enhanced Shakespearean Corpus: Verse (hereafter, *ESC:Verse*)

All elements are openly available under a Creative Commons licence (at present via Lancaster University's server and in future as full downloads).[1] *ESC:Comp* and *ESC:EEBO* are fully described and discussed in two earlier papers (respectively, Demmen 2020 and Murphy 2019); this paper focuses on *ESC:Folio*, with additional notes on *ESC:Quartos* and *ESC:Verse*, whose markup and annotation are largely similar to *ESC:Folio*. Despite the chronology of publication, *ESC: Folio* was the first component designed and created, and the one in the context of which many of the principles and procedures subsequently used for other components were developed. Importantly, the issues in focus do not pertain solely to the language of Shakespeare, but are relevant to many kinds of corpus construction and exploitation, especially those involving historical data.

The paper is organised as follows. We begin in Section 2 by presenting the composition of *ESC:Folio*, *ESC:Quartos* and *ESC:Verse*. After considering the conceptual issue of what kind of historical entity a corpus of Shakespeare's drama can or should represent, we move on to detail the textual composition of the corpora. *ESC:Folio*, in particular, is already 'enhanced' at this stage by the inclusion of plays not present in the First Folio edition from which most of the corpus drawn. The remainder of the paper proceeds in order through the other forms of enhancement which we applied to the corpora. Sections 3 and 4.1 recount, respectively, how we sourced the base electronic versions of the texts from *Internet Shakespeare Editions* and how we then reworked and reorganised the XML markup to optimise it for corpus analysis – including, but not limited to, the needs of the *Encyclopedia* project. The account of the corpus markup in Section 4 is then completed, in 4.2, by a discussion of the extensive character metadata that is linked to the markup of spoken turns in the plays. The final family of enhancements we explore, in Section 5, are those related to the different kinds of analytic annotation applied to the corpus. We address spelling regularisation (5.1); part-of-speech tagging (5.2); and lemmatisation and semantic tagging (5.3). For each type of annotation, we explain the benefits it brings to the corpus, the problems we encountered when implementing it in the Enhanced Shakespearean Corpus, and the approaches we took to addressing those problems.

2 Composition of ESC: Folio and secondary components

The goals of the *Encyclopedia* project necessitated the definition of a coherent object of study from which a single, stable corpus could be built as the primary basis of analysis. Many editions exist of Shakespeare's plays. During his life, they were published individually (in quarto format), with collected editions (in folio format) following shortly after his death. Over the centuries, and of course continuing to the present day, many further editions have been produced. The path of least resistance for the project would have been to base our corpus on such a modern edition. This would allow us to benefit from the critical textual analysis such editions often incorporate (by adding it to the corpus as annotation). There are, however, three reasons why we did not take this path. First, modern editions typically do not, in fact, represent the kind of coherent entity that we needed. Rather, they are often collations of multiple extant versions: the First Folio of 1623 (F1)[2] and the quarto texts (for those plays for which quartos exist). With a variety of audiences in mind (but not usually linguists), editors combine versions, using preferred readings from each so as to amend perceived errors in order to reconstruct, or perhaps reimagine, a hypothetical or 'best-guess' historical text – as the footnotes to any modern edition reveal – which is thus not actually a historical text that can act as an anchor point. Second, modern editors often strip out or normalise into uniformity the very things whose nature or variability might interest a linguist, such as the Early Modern English inflections of verbs and nouns. This reflects the non-linguistic priorities of such editions, for instance, to create a readable or performable script.

Finally, editorial practice, especially regarding the micro-detail of the language, is not always consistent. How compounds are treated is a case in point. For example, we discovered that the word *hourglass* was rendered as *hourglass*, *hour-glass* and *hour glass* in each of three different modern editions of Shakespeare's texts, different editors altering the original seventeenth-century rendering(s) (itself likely to present spelling variation beyond just the spacing) in different ways. For the general reader, of course, how compounds are spaced out is hardly an impediment. But for any statistical corpus-based analysis, acute problems are raised: an instance of an open compound adds one to the frequency of *each component*, whereas an instance of a hyphenated or fully-closed compound adds one to the frequency of *the compound* as a separate word-type.

These considerations dictated that we should base the corpus on a specific particular edition from Shakespeare's time, that is, a genuine historical entity; and, moreover, relying on an unmodified, original spelling transcription of that edition allows us to control consistency in the treatment of matters like com-

pounds, other spelling variation, and so on. This being the case, the most obvious version to choose was F1: as the earliest publication of a collected edition, its texts do collectively constitute a coherent object of study. This is not to say that no controversies attend the text of F1. Scholars have recognised the presence of other hands in some plays within F1, despite the attribution in its title: *Mr William Shakespeare Comedies, Histories and Tragedies*. For example, *The New Oxford Shakespeare*'s version of the "complete works" (edited by Taylor et al. 2016) lists Christopher Marlowe, Thomas Nashe, George Peele, Thomas Heywood, Ben Jonson, George Wilkins, Thomas Middleton and John Fletcher as co-authors. Collaboration works both ways, of course, and there are many claims of Shakespeare's hand in plays *not* in F1; famously, one manuscript page of *The History of Thomas More* is believed to be an example of Shakespeare's handwriting.

In developing our corpus, it would have been possible to exclude portions of F1 widely deemed to be by writers other than Shakespeare, and to add (portions of) other plays considered to be by Shakespeare. We did not, however, do this. First, this approach would have ended the correspondence of the corpus with any coherent historical object. Second, attempting to include all the plays with scenes or fragments now attributed to Shakespeare (such as *Arden of Faversham*, *Double Falsehood* and *Edward III*) would have embroiled the project in (interminable) debates about authorship attribution before even getting off the ground. Rather, we stuck largely to F1 as the basis for the corpus. We added to this only the Q1 editions of two plays not in F1 that nevertheless have a relatively longstanding scholarly acceptance into the Shakespeare canon: *Pericles* and *The Two Noble Kinsmen*, believed to be collaborations with George Wilkins and John Fletcher respectively. The corpus name, *First Folio Plus*, is motivated by this: it consists of F1 *plus* two more plays. Adding them means that our corpus moves slightly away from being an exactly authentic reproduction of a specific seventeenth-century physical publication (that is, F1). But the overall set of 38 plays *does* correspond to what a hypothetical Shakespeare completist 'fan' of circa 1625 would have been able to access in print. Similarly, it corresponds closely to the collection of the texts that the general public today (circa 2025!) associates with Shakespeare as a dramatist. We consider this a justifiable and defensible line to draw around our primary dataset.

To be clear, this decision does not constitute a claim on our part that F1 is in some sense more validly 'Shakespearean' than the Quartos or modern critical editions. We have some sympathy with recent comments made by Greg Doran, artistic director of the Royal Shakespeare Company, on whether or not Shakespeare wrote his plays: [3]

> I don't care – ultimately we have this fantastic body of work – it doesn't matter who he, she or they were, in a way, we've got them – we can debate endlessly about who the writer was

For Doran, the texts are important as a catalyst for performance; for us, they constitute, in their form in F1, an empirically grounded basis for accounts of a particular type of language from a particular socio-historical context. The status of F1 as a coherent historical artefact is, we would argue, wholly independent of any authorship question. As Doran notes, whether or not Shakespeare wrote any particular work (in part or in full) is not provable, although the work of authorship attribution scholars is far from irrelevant to the extent that it assists understanding of relationship between the text in focus and William Shakespeare the *person* (as opposed to William Shakespeare, the body of literary works).

Table 1 lists the plays that constitute *ESC:Folio*, plus additional information accessible within the corpus's text-level metadata. The plays are listed by their most usual modern short title (the original editions had much longer titles). The abbreviated titles, which follow those used by the *Arden Shakespeare*, are used as text ID codes in the corpus markup and metadata.

Table 1: The plays constituting *ESC:Folio*

Play (short title)	Abbreviation	Genre: [T]ragedy, [C]omedy, [H]istory	Date of first publication	Date range for first production	Approximate date of first production
Titus Andronicus	Tit	T	1594	1590–1592	1592
Romeo and Juliet	RJ	T	1597	1594–1595	1595
Julius Caesar	JC	T	1623	1598–1599	1599
Hamlet	Ham	T	1603	1600–1601	1601
Troilus and Cressida	TC	T*	1609	1602–1603	1602
Othello	Oth	T	1622	1603–1604	1604
King Lear	KL	T	1608	1605–1606	1605
Timon of Athens	Tim	T	1623	1605–1606	1605
Antony and Cleopatra	AC	T	1623	1606–1608	1606
Macbeth	Mac	T	1623	1606	1606
Coriolanus	Cor	T	1623	1608	1608
Henry VI, Part 2	2H6	H	1594	1590–1591	1591
Henry VI, Part 3	3H6	H	1595	1591	1591

Henry VI, Part 1	1H6	H	1623	1590–1592	1592
Richard III	R3	H	1597	1591–1593	1592
Richard II	R2	H	1597	1595	1595
King John	KJ	H	1623	1596	1596
Henry IV, Part 1	1H4	H	1598	1596–1597	1597
Henry IV, Part 2	2H4	H	1600	1597–1598	1597
Henry V	H5	H	1600	1598–1599	1599
Henry VIII	H8	H	1623	1613	1613
Much Ado about Nothing	MA	C	1600	1598	1598
Two Gentlemen of Verona	TGV	C	1623	1590–1591	1590
The Taming of the Shrew	TS	C	1623	1590–1604	1592
The Comedy of Errors	CE	C	1623	1590–1594	1594
Love's Labour's Lost	LLL	C	1598	1594–1595	1595
A Midsummer Night's Dream	MND	C	1600	1595–1596	1595
The Merchant of Venice	MV	C	1600	1596–1598	1596
The Merry Wives of Windsor	MW	C	1602	1597–1598	1597
As You Like It	AYL	C	1623	1598–1600	1599
Twelfth Night	TN	C	1623	1601–1602	1601
All's Well that Ends Well	AW	C*	1623	1603–1604	1603
Measure for Measure	MM	C*	1623	1603–1604	1603
Pericles (Q1)	Per	C	1609	1606–1608	1608
The Winter's Tale	WT	C	1623	1609–1611	1609
Cymbeline	Cym	C	1623	1608–1611	1610
The Tempest	Tem	C	1623	1611	1611
The Two Noble Kinsmen (Q1)	TNK	C	1634	1613–1614	1613

Classifying Shakespeare's plays by genre is not uncontroversial. Table 1 uses the designations in F1 itself to identify each play as a tragedy, comedy or history (with the exception of *Cymbeline*, which in F1 is a tragedy but is classed in

ESC:Folio as a comedy, according to convention in modern editions). Despite their widespread familiarity, however, the categories of tragedy, comedy and history are slippery notions, and some of the F1 designations are puzzling. A case in point is the so-called 'problem plays', conventionally taken to be *All's Well That Ends Well*, *Measure for Measure* and *Troilus and Cressida*, which combine elements of comedy and tragedy (marked as T* or C* in Table 1). To give users of *ESC:Folio* the option of tracking these three plays as a group, allowing any specific characteristics to be revealed, an additional metadata field was added in which they are assigned a distinct class from the other three genres (coded P).

The probable dates of first publication are supported by fairly good evidence.[4] The evidence for dates of first stage production is often rather thinner. Thus, two fields are provided: the range in which the first production probably occurred, and an approximate specific date – a 'best guess' year which permits roughly-correct chronological sorting. All dates are drawn from the authoritative *Database of Early English Playbooks (DEEP)* (http://deep.sas.upenn.edu/; see Farmer and Lesser 2007).

Whilst *ESC:Folio*, as noted, contains the two quarto plays not in F1, we did not wish to entirely neglect quarto editions of plays that *are* in F1. These earlier editions often differ from F1. For some plays, the differences are minor, involving just a few words. But in other cases differences are rather more substantial, even involving whole scenes; modern editors naturally vary in their preferences for F1 readings or quarto readings. Creating the *ESC:Quartos* corpus alongside, but separate to, *ESC:Folio* made it possible for analyses based on F1 texts to be checked against what might emerge from the quartos. *ESC:Quartos* collects the 22 quarto editions of text-critical significance (that is, *not* late-published quartos known to be derivatives of other extant editions). All are Q1 or Q2 editions, that is, the first or second known publications of the plays they represent (Q0 of *Henry IV, Part 1* is a partially-surviving text). Table 2 lists the texts and their publication dates.

Table 2: The plays constituting ESC:Quartos

Play	Edition	Date of publication	Text ID
Henry VI, Part 2	Q1	1594	2H6_Q1
Titus Andronicus	Q1	1594	Tit_Q1
Henry VI, Part 3	O1	1595	3H6_O1
Richard II	Q1	1597	R2_Q1
Richard III	Q1	1597	R3_Q1
Romeo and Juliet	Q1	1597	RJ_Q1
	Q2	1599	RJ_Q2
Henry IV, Part 1	Q0	1598	1H4_Q0
	Q1	1598	1H4_Q1
Love's Labour's Lost	Q1	1598	LLL_Q1
Henry IV, Part 2	Q1	1600	2H4_Q1
Henry V	Q1	1600	H5_Q1
The Merchant of Venice	Q1	1600	MV_Q1
A Midsummer Night's Dream	Q1	1600	MND_Q1
Much Ado about Nothing	Q1	1600	MA_Q1
The Merry Wives of Windsor	Q1	1602	MW_Q1
Hamlet	Q1	1603	Ham_Q1
	Q2	1604	Ham_Q2
King Lear	Q1	1608	KL_Q1
	Q2	1619	KL_Q2
Troilus and Cressida	Q1	1609	TC_Q1
Othello	Q1	1622	Oth_Q1

One play published in quarto in 1594 which we exclude from *ESC:Quartos* is *The Taming of a Shrew* (note "a" not "the"), a text of debated Shakespearean authorship whose relationship to the similar, definite-article-named play in F1 is likewise a matter of debate (for introductions to this issue, see Kelly n.d.; Schafer n.d.). This omission was an extension of the principle that led us to omit *Edward III* and the like from *ESC:Folio*. On the same basis, the first publication date we record for *The Taming of the Shrew* in *ESC:Folio*'s metadata is 1623, not 1594.

Finally, let us turn to *ESC:Verse*. Segregating the poetry from the plays, while still having it available for analysis, was critical, because Shakespeare's poems are especially dense in technical problems for corpus-based analysis, even relative to the plays. Some issues are minor, e.g. differences in textual structure. But critically, many of the automated corpus annotation processes we applied to *ESC:Folio* (see Section 4 below), most notably part-of-speech tagging, are liable to suffer from reduced accuracy in the poetry. Our tagger is partially probabilistic, using frequency data on (a) what part(s)-of-speech each word normally has, and (b) what sequences of word classes are normal for English. But these are precisely the norms that poetry often violates for effect. A higher error rate in the tagging of the poetry is acceptable, because allowances may be made for it, when it is a separate dataset; it would not be acceptable were the poems treated together with the plays. This consideration motivated *ESC:Verse*'s creation as a separate corpus, albeit a small one. Its content (see Table 3) requires little comment. *The Passionate Pilgrim* is a contested text, an anthology of twenty poems of which five are Shakespeare poems also found in other texts (the *Sonnets* and *Love's Labour's Lost*), and the remaining fifteen are either known not to be by Shakespeare or of unknown authorship; in each of the latter cases, there exists, naturally, debate about which if any are more or less likely to be by Shakespeare (Duncan-Jones and Woudhuysen 2007:84–91). Following our principle of limiting our engagement with authorship debates, we make no attempt in *ESC:Verse* to separate what is by Shakespeare from what is not, and simply present the Octavo text. *The Phoenix and the Turtle*, on the other hand, is a single three-page Shakespeare poem published as part of a multi-author appendix to a much longer work, *Love's Martyr* by Robert Chester (Duncan-Jones and Woudhuysen 2007: 91). In this case, only the approximately 400 words by Shakespeare are included in *ESC:Verse*. Finally, following tradition in modern editions, *A Lover's Complaint* is treated in *ESC:Verse* as a separate text to the *Sonnets*, although they were published as a single Quarto volume. Table 4 presents overview statistics for all three ESC components.

Table 3: The poems constituting *ESC:Verse*

Work	Edition	Date of publication	Text ID
Venus and Adonis	Q1	1593	VA
The Rape of Lucrece	Q1	1594	Luc
The Passionate Pilgrim	O2	1599	PP
The Phoenix and the Turtle	Q1	1601	PhT
The Sonnets	Q1	1609	Son
A Lover's Complaint	Q1	1609	LC

Table 4: Overview statistics

Corpus	Texts	Tokens (nearest thousand)
ESC:Folio	38	1,039,000
ESC:Quartos	22	536,000
ESC:Verse	6	56,000

3 Source transcriptions for ESC: Internet Shakespeare Editions

The original electronic files for *ESC:Folio*, *ESC:Quartos* and *ESC:Verse* were generously provided to the *Encyclopedia* project by Internet Shakespeare Editions (ISE),[5] with the kind permission of the University of Victoria. As their website recounts, ISE "has from the beginning had the highest standards of academic development in mind – and these standards are overseen and maintained by a distinguished editorial board from around the world".[6] Crucially, ISE's versions of the F1 and quarto texts had not undergone any of the modifications associated with modern editions which, for reasons discussed previously, would have been problematic for our purposes. Specifically, ISE state that their "old-spelling versions are diplomatic transcriptions and do not amend any errors present in the original text".[7] Diplomatic transcriptions are unedited, uncorrected 'warts-and-all' transcriptions; this entirely suited our aim of building from the foundation of an original, untampered representation of F1.

The ISE transcriptions were initially based on a facsimile edition of F1 (Hinman 1968), but then also electronically checked against transcriptions held in the Oxford Text Archive, with appropriate corrections made at that point. They may be seen as embodying the closest that it is possible to approach a faithful representation of F1 as digital text (as opposed to digital *images*). Human fallibility, and the limits to the flexibility of machine-readable text storage, are such

that a *completely* faithful transcription of an extensive historical document is more of an ideal to be pursued than an achievable reality. Given ISE's prior extensive pursuit of the ideal, we considered it superfluous to ourselves attempt any systematic proofreading of the transcriptions against page images of F1. But, bearing in mind the possibility of some small number of errors having evaded ISE's extensive quality control measures, we did investigate, and if necessary correct, any oddities in the transcription that we noticed while working on the markup and annotation of the corpora. Typically, those oddities concerned devices that early modern English printers sometimes deployed to save space, such as the use of vowel elisions, superscript and words split at the ends of lines. These tend to be the very things that computer optical character recognition programs struggle with or even the human eye fails to see.

4 Structural markup and metadata

4.1 XML markup

The ISE transcripts are encoded as XML (Bray et al. 2008). XML is an extraordinarily flexible system of markup, as each application of XML defines its own elements, attributes, and rules for their use. The XML schema used by ISE is oriented towards preserving the highest possible degree of detail of the visual appearance of the original pages of F1. For instance, the beginning of the play *Henry IV, Part 1* is encoded as follows (with line breaks added for readability):

```
<l align="center"> <ms t="tln" n="2"/><sd t="entrance"><i> Enter
the King, Lord Iohn of Lanca<lig unicode="ft"><typeform
set="f">s</typeform>t</lig>er, Earle</i></sd></l>
<l align="center"> <sd t="entrance"><ms t="tln" n="3"/><i>of
We<lig unicode="ft"><typeform set="f">s</typeform>t</lig>merland,
with others</i>.</sd></l>
<l> </l>
<mode t="verse"/>
<l align="center"> <ms t="tln" n="4"/><s k="1"><sp
norm="King"><i>King</i>.</sp></s></l>
<l> <s k="1"><ms t="tln" n="5"/><ornament drop="4">S</ornament>O
<lig><typeform set="f">s</typeform>h</lig>aken as we are,
<typeform set="f">s</typeform>o wan with care,</s></l>
<l> <s k="1"><ms t="tln" n="6"/>Finde we a time for frighted Peace
to pant,</s></l>
```

An account of this complex schema is beyond the scope of this paper. It suffices to note that the elements used to structure the document are based primarily on layout; most notably, the primary division is by line, that is < l> tags (marked as

centre-aligned where necessary). Beyond that, highly complex markup is used to record the original form of the printed characters. The <typeform> tag records every case of a long-S character being encoded as *s*; the <lig> tag records every instance of two letters represented by a ligature. But there are also interpretive tags. <sd> encloses a stretch of stage direction, <sp> encloses a speaker label, with a normalised name for the character (e.g. norm="King"), and <s> encloses actual speech, with the k= "*N*" attribute indicating what turn each <s> belongs to.

Much of this detail was superfluous for our purposes. A corpus text – as opposed to a digital edition or archive – is a linguistically analysable representation of an original communicative event (here, the printing of the play within F1). From this perspective, it is more important to record that the stage direction contains an instance of the word *Lancaster* than to record that the *-st-* was printed as a long-S in ligature with *t* – especially when recording the latter makes the word discontinuous. On the other hand, information on spoken lines, stage directions, character labels, and act/scene boundaries (not exemplified in the extract above) were clearly of relevance.

We therefore reformatted the XML so as to jettison *all* the presentational elements discussed so far, as well as tags indicating:

- italic text
- the two-column layout of F1
- the running heads at the top of each page
- the printing in the extreme lower right of each page of the first word from the next page (common in the period)
- the wrapping of long lines to unused space on right edge of a preceding or following line (a spaced-saving device common in collected editions of Shakespeare to this day).

We also changed the names and exact usage rules of many of the tags we did retain. Our goal was a structure along the lines described in Hardie's (2014) recommendation for the use of XML markup in corpus linguistics. Most of the changes were made programmatically, by a combination of (a) parsing the XML document tree, manipulating its structure, and then writing the modified tree back to file; and (b) performing global search-and-replace using regular expressions. The complexity of the ISE's XML schema, and the 'warts' of F1, meant that the automated mapping had to be supplemented by manual intervention to finalise the XML. The end point of this was a structure in which the passage from the ISE version of *Henry IV, Part 1* given above appears as follows:

```
<stage> Enter the King, Lord <normalised orig="Iohn"
auto="false">John</normalised> of Lancaster,
<normalised orig="Earle" auto="false">Earl</normalised>
<lb/> of <normalised orig="Westmerland"
auto="false">Westmorland</normalised>, with others.
</stage>
<lb/>
<u who="1H4_King" label="King."> So shaken as we are, so
wan with care,<lb/> <normalised orig="Finde"
auto="false">Find</normalised> we a time for frighted
Peace to pant,<lb/>
```

These four lines are a fairly representative sample of the XML of *ESC:Folio* overall, and exemplify many of the elements and attributes whose usage the remainder of this section will lay out.

The primary structural units are <text>, <act> and <scene>. Each of these has an ID code that is globally unique in the corpus as its *id* attribute. Each also has a *title* attribute, which preserves that unit's heading (if any). Once moved into the XML, these titles were no longer treated as parts of the running text of the plays. The sequential numbering of acts and scenes is preserved as *n* attributes; IDs of acts and scenes are programmatically generated from the play ID and these numbers. The following examples are from the beginning of *Othello*; the closing tags for the elements, that is </scene>, </act>, and </text>, are not shown, as they come much later in the file.

```
<text title="THE TRAGEDIE OF Othello, the Moore of
Venice." id="Oth">
<act n="1" title="Actus Primus" id="Oth_1">
<scene n="1" type="scene" title="Scoena Prima"
id="Oth_1_1">
```

The final attribute of note is the *type* attribute of <scene>. This has three possible values: *scene*, *prologue*, and *epilogue*. These values were inserted manually. A final structural element, <endMatter>, encloses parts of plays after the end of their last scene. Frequently, a play's end matter is no more than the single word *FINIS*. In some cases, however, it is more extensive. The purpose of the <endMatter> tags is largely to allow these parts of the corpus to be excluded from computer analysis whenever necessary.

Within scenes, the main elements are <stage>, which encloses stage directions; <u>, short for utterance, which represents a single turn of character

speech; and <lb/> for line breaks. All are exemplified in the *Henry IV, Part 1* excerpt above. While they are mapped from the ISE XML elements for these same concepts, we made numerous tweaks. First, moving from <l>…</l> around each line, to a unitary <lb/> indicating the point location of a line break, means that lines are no longer the principal structural unit: instead of <stage> and <u> being inside lines, line breaks are inside (or between) instances of <stage> and <u>. This meant that adjacent stage directions could be merged to reflect the conceptual organisation more directly. With other tags omitted to make the transformation clear:

```
ISE XML:
<l><sd>Enter the King, Lord Iohn of Lancaster, Earle
</sd></l>
<l><sd>of Westmerland, with others.</sd></l>

ESC:Folio XML:
<stage> Enter the King, Lord John of Lancaster,
<normalised orig="Earle" auto="false">Earl <lb/> of
Westmorland</normalised>, with others.</stage>
```

Similarly entire sequences of ISE <s> elements were merged together to generate unbroken <u> elements. Where a line break occurs mid-word, in *ESC:Folio* the <lb/> tag is placed after the complete word, a minor misrepresentation which preserves word token continuity.

Stage directions, of course, occur within stretches of character dialogue on occasions. To avoid any text being marked as both speech and stage direction, we added utterance boundaries before and after any such embedded stage direction: the <u> tag ends, the <stage> element wraps the direction, and then a new <u> begins to continue the speech. Inducing these utterance breaks was another process which was automated but required substantial subsequent manual adjustments to deal with edge cases and errors in the automated conversion.

The <u> tags are crucial for our work, as they allow the language of different characters to be tracked. The following full utterance example is from *Othello* 1.3:

```
<u who="Oth_DukeOfVenice" label="Duke."> There's no
composition in this <normalised orig="Newes"
auto="false">News</normalised>,<lb/> That
<normalised orig="giues" auto="false">gives
</normalised> them <normalised orig="Credite"
auto="false">Credit</normalised>.<lb/></u>
```

The speaker label that precedes the actual words of this short speech, "Duke.", has been moved into the *label* attribute so that – as with the act and scene titles – it is no longer part of the principal running text. In cases where no speaker label is present (e.g. when a turn ends and then resumes around an embedded stage direction) the *label* attribute is present, but empty.

The *who* attribute requires a little more comment. The normalised names coded as *norm* attributes in the ISE XML (see the *Henry IV, Part 1* example above) accurately unify the turns spoken by characters who are labelled in multiple ways in F1. In *Othello*, the Duke of Venice's turns are consistently labelled "Duke." But Othello's turns often have abbreviated labels: "Oth. / Othe. / Othel." as well as "Othello", so that only the normalised name associates them. While we needed to retain this functionality, for our purposes the character identifiers needed to be globally unique in the corpus. "Othello" and "Duke of Venice" are globally unique, but other normalised names are not. For instance, six plays[8] have a character with the normalised name "Duke". Moreover, our use of character identifiers in the speaker metadata of the corpus (see 4.2 below) meant that shorter codes with no spaces were preferable to phrases like "Duke of Venice", to support their use as database keys. Thus, another automatic mapping was implemented to generate unique IDs from the normalised names, by removing spaces, dropping definite articles, and prepending the play abbreviation to distinguish among the six characters with normalised name "Duke"; these unique IDs then *replaced* the normalised names in the actual markup. The transformation is as follows:

```
ISE XML:          norm="Duke of Venice"
ESC:Folio XML:    who="Oth_DukeOfVenice"
```

Finally, within the actual text of stage directions and utterances, two further elements are used. The <normalised> element stores spelling regularisation. The regular form appears between the opening and closing tags; the source spelling is captured by the *orig* attribute; the *auto* attribute records whether the change was made automatically (*true*) or manually (*false*). Several examples of <normalised> may be found in the excerpts already presented. The process by which the spelling regularisation was introduced using the VARD2 software will be discussed in Section 5.2. The <foreign> element encloses words or stretches of words that are deemed to be non-English; its *lang* attribute stores a label for the language of the words thus enclosed. These tags were added manually during spelling regularisation using VARD2. This speech from *The Two Noble Kinsmen* illustrates the use of <foreign>:

```
<u who="TNK_Schoolmaster" label="Sch,"> <normalised
orig="Goe" auto="false">Go</normalised> take her,
<normalised orig="aud" auto="false">and</normalised>
fluently <normalised orig="perswade" auto="false">
persuade</normalised> her to a peace:<lb/>
<foreign lang="Latin">Et opus exegi, quod nec
<normalised orig="Iouis" auto="false">Jovis
</normalised> ira, nec ignis</foreign>.<lb/> Strike up,
and <normalised orig="leade" auto="false">lead
</normalised> her in.<lb/></u>
```

We extended the use of <foreign> from actual non-English words, as above, to words whose form in the plays as printed indicates that they were to be delivered with the pronunciation of some given dialect. A famous example is the regular substitution of the letter *b* with the letter *p* in speeches for Welsh characters (e.g. in *Henry V*, Fluellen's *pig* for *big*). Such 'eye-dialect' is not, by its nature, likely to represent the accents of English of the early modern period in more than the broadest, most stereotypical strokes. For that reason, we differentiated only five categories of 'dialect' according to region: *English*, *Irish*, *Welsh*, *Scottish* and *Other*. *Dialect – Other* was used for English words spelt so as to indicate the pronunciation of a non-native speaker, such as French-accented English words (e.g. *dat* for *that*; many examples can be found in *Henry V*, 5.2). The 'dialect' spellings were subsequently regularised (see 5.1 below) so that corpus queries for *big* will indeed retrieve Fluellen's *pig*, with the <foreign> and <normalised> tags constituting the record of that change.

The different values in the corpus for the *lang* attribute are listed in Table 5.

Table 5: Foreign languages and dialects marked up in ESC:Folio

Language or dialect	Tokens
Dialect – English	43
Dialect – Irish	18
Dialect – Scottish	20
Dialect – Welsh	201
Dialect – Other	326
French	1,049
Italian	72
Latin	1,336

Middle English	6
Spanish	45
Other Language*	191

**Other Language includes 135 examples of the word signior, some other forms which like signior are ambiguous between Italian and Spanish, and some nonsense words.*

The high level of French words in *ESC:Folio* is largely due to *Henry V*, which contains scenes partly or mostly in French – the dubious bilingualism of King Henry of England and Princess Catherine of France being both a minor plot point and a pretext for some salacious puns. Meanwhile, the conventional use of *exeunt*, *manet*, *finis*, and *solus* in stage directions and end matter accounts for 798 of the Latin tokens. The listing of Middle English as a distinct language may be surprising. The six words in question are archaisms which exhibit inflectional affixes already extinct in Shakespeare's day (*yclad, ycleped* (twice), *killen*, *yslacked*, *yravished*). Classing these as foreign meant that we did not need to force them into the framework of Modern English grammar at the part-of-speech tagging stage (see 5.2); other possible treatments, such as tagging them as if they were the equivalent Modern English forms, were found to lead to unwanted inconsistencies in lemmatisation.

Unlike conventional *exeunt* and friends, the conventional stage direction *exit* was not marked up as a foreign, Latin word – since we deemed to *exit* to be an English loanword, rather than a foreign word used in a kind of code-switching. On occasion, determining some word's 'foreign' status in this way was not straightforward; was it really foreign or had it become naturalised as a loanword? In such cases, we made a judgement based on the word's contexts of use, not just in Shakespeare but also more broadly in Early Modern English.[9] The easiest calls were words which exhibited English inflections (definite loanwords) or that appeared within whole phrases in the donor language, as in the *Two Noble Kinsmen* example above (definite foreign words). Exhibiting donor-language inflection was less decisive, since loanwords (including, as it happens, *exit*) may preserve donor-language affixes as part of the new English root. In the (many) cases where these considerations did not decide the matter, we depended on whether it generally appeared in the context of texts or speakers of the language in question. The French-origin *Monsieur* is a case in point. For such a pre-nominal title, neither inflection nor syntax are helpful guides. But exploring relevant examples of usage reveals that it was in Shakespeare's period used most

frequently in contexts of interaction with French people. The determination was therefore that it was a foreign word, and it was marked up as such.

Before leaving the subject of XML markup, a few observations remain to be made regarding how the ESC components other than *ESC:Folio* utilise or modify the system explained so far. *ESC:Comp* (Demmen 2020) uses the *ESC:Folio* system without change, except that it required a <frontMatter> element to mirror <endMatter>. *ESC:Quartos* deviates from *ESC:Folio* and *ESC:Comp* only in the absence of the *who* attribute for utterances. This is because *ESC:Quartos* lacks the linked speaker metadata that necessitates ID codes for characters in *ESC:Folio* and *ESC:Comp*. Instead, in *ESC:Quartos* the ISE normalised names are retained in attribute *norm* on the <u> element. The XML of *ESC:EEBO* (Murphy 2019) is radically simpler. Only <text>, its *id* attribute, and <p> tags demarcating paragraphs are present, <p> being the only element preserved from the original EEBO-TCP XML files. Finally, in *ESC:Verse*, <text>, <lb/> and <endMatter> are used as in *ESC:Folio*, but each text is also subdivided. In the *Sonnets*, the subdivisions are <poem> elements, with a *name* attribute with values *Sonnet 1*, *Sonnet 2*, etc., derived from ISE. The other texts consist of one long poem each and so are divided into <stanza> elements.

4.2 Character metadata

Not everybody speaks alike; language varies in interesting ways across society. Sociolinguists have investigated the different ways in which male talk and female talk are constructed, and how people higher up a social hierarchy differ from those lower down. There are not, in fact, distinct speakers behind the different characters in the Shakespeare plays, all their speech being the product of the playwright(s). But the social categories of the speakers in the fictional worlds of Shakespeare's plays are still of interest, in that any correlation of linguistic features with these categories will shed light on the socio-literary-discoursal construction of *real* historical (groups of) people through their representation in drama by Shakespeare and others of the period (cf. our *ESC:Comp* corpus, Demmen 2020).

The unique identifiers assigned to each utterance in *ESC:Folio* are references to a metadata table for the character who speaks that turn. The 'speaker'[10] metadata includes the standardised character name, as in the original ISE files (see 4.1) with only minor tweaks, as well as a list of the abbreviations used to label that character's speech in F1. Moreover, the play to which the character belongs is listed (characters appearing in more than one play, such as Richard of Gloucester/Richard III, have separate identities per play). Incidental minor characters referred to only by profession or status rather than name – messengers,

servants, thieves, soldiers, lords, senators, etc. – are not distinguished if the text does not distinguish them (e.g. by numbering them).[11] As common in drama, often in F1 lines are assigned to two or more speakers, or to a group such as *Lords*, or to *All*. These groups have distinct identifiers to the characters who make them up and, often, distinct values for other metadata properties; to distinguish groups from individuals, the *number* of speakers represented by each ID is recorded. In addition to this, a number of social categories are recorded. The full list of metadata fields is given in Table 6.

Table 6: 'Speaker' metadata for characters in F1

Metadata feature	Possible values
Speaker ID tag	Unique code for every character
Normalised name	Standardised (modern spelling) character name
Label list	Pipe-delimited list, e.g. "Ro.|Rom.|Rome.|Romeo." for Romeo in RJ
Play	The text ID codes from Table 1, e.g. LLL, MW, TNK
Number of speaker(s)	Singular (s) or multiple (m)
Sex of speaker	Male (m), female (f), assumed male (am), assumed female (af), problematic (p), assumed problematic (ap)
Status/social rank of speaker	Monarch (0), nobility (1), gentry (2), professional (3), other middling groups (4), ordinary commoners (5), lowest groups (6), supernatural beings (7), assumed status categories (a0 through a7), problematic (p), assumed problematic (ap)

Categorising characters as male or female is relatively straightforward, with the exception of assumed identities, to which we return below; for mixed-sex group speaker IDs, we record *problematic* (p). Other than sex, age is perhaps the most common metadata attribute in spoken corpora, but in the Shakespeare plays, age information is deducible only for some characters, and in some cases only vaguely;[12] thus we made no attempt to include age groups.

However, characters' social status (or class, or rank) *can* usually be worked out, and is of major importance. In theory the gradations of class could be very fine indeed, but for the practical purpose of grouping for quantitative analysis, something more coarse-grained is required. The nine-category scheme we use draws upon the approach developed by Archer and Culpeper (2003). It is designed to reflect the pre-industrialised nature of early modern society, based on a review of relevant historical work.[13] The categories are also intended to

reflect the way in which early modern commentators wrote about status. So, for instance, Sir Thomas Smith (1583: 20) pointed out that men (*sic*) could be divided into "foure sortes" in the latter sixteenth century, namely, gentlemen (in which he included nobility), citizens, yeoman artificers and labourers. Thanks to the work of historians, we know that the population of England grew from around three million to around four million during Elizabeth I's reign (1558–1603); estimates for the number of gentry vary, but the nobility and upper gentry are thought to constitute a couple of percent of the population – the vast majority of the population belonged to classes below the gentry (see the summary of historical studies in Nevalainen and Raumolin-Brunberg 2003: 32–38). As such, we have been careful to adopt a categorisation scheme which can distinguish gentry from professionals and other middling groups, as well as distinguish ordinary commoners from the lowest groups. Our social status categories are shown in Table 7, along with the numbers used to code them, a brief explanation and prototypical examples. Six categories (Gentry, Professional, Other Middling Groups, Ordinary Commoners and Lowest Groups) are adopted unchanged from Archer and Culpeper's (2003) system. Archer and Culpeper's Nobility category has been split to code Monarchy separately from Nobility. Monarchy merits a separate social class on the basis that, for many of Shakespeare's contemporaries, God alone was superior to the sovereign; everyone else was a subject of the Queen or King. The eighth numbered category, Supernatural Beings, accounts for the forty-plus ghosts, god, fairies, etc. in the plays. Group speaker IDs of mixed social status are assigned *problematic*, as are characters whose status changes during the play – except for the common and straightforward case of a noble becoming king or queen; characters who undergo that transition are simply placed in category 0.

Table 7: Social status categories

Social category	Code	Description
Monarchy	0	Rulers of subjects. Prototypical examples – King, Queen, Majesty. Less prototypical examples – Duke, Prince (only where sovereign e.g. over a city state).
Nobility	1	Those with inherited or conferred titles that would allow them to sit in the House of Lords, including the 'Lords spiritual'; or equivalent ranks in other countries/cultures. Prototypical examples: Duke, Marquess, Earl, Viscount, Baron, Archbishop, Bishop.

Gentry	2	Upper Clergy and non-hereditary knights not able to sit in the House of Lords, people entitled to carry arms and/or recognised as having the (legitimate) capacity to govern, and those able to append the title *esquire* (Esq.) to their name legitimately. Likely to be of a certain income, substantially above £2,000 per annum. Prototypical examples: Knight, Sir, Major General.
Professional	3	Those practising high-level skills, including civil servants, teachers, army and naval officers and members of the 'learned professions', or the 'three great professions' of Law, Medicine and the Church. Prototypical examples: clergyman, lawyer, medic, schoolteacher.
Other Middling Groups	4	Those directly involved in trade and commerce, whose focus is upon production/distribution as opposed to service and whose income is likely to have been between £50–£2,000. Prototypical examples: manufacturer, wholesaler, retailer, merchant, money-lender, skilled craftsman, financier.
Ordinary Commoners	5	Those who laboured on someone else's materials or in someone else's fields, household or manufactory, and whose income is likely to have been less than £50 per annum. Prototypical examples: yeoman, poor husbandman, wage labourer, apprentice.
Lowest Groups	6	Prototypical examples: common seaman, servant, cottager, pauper, common soldier, vagrant.
Supernatural Beings	7	Prototypical examples: ghost, fairy, god, sprite, apparition.
Problematic	p	Those whose status was uncertain at the time, e.g. actor; 'multiple' character identities where the group members have mixed status; characters who undergo a significant change in status during the play.

Evidence of descriptive features, as listed in Table 7, is not always available (e.g. we are very rarely given precise information about a character's income). Obviously, judgement calls were required in some cases. For instance, for the Roman plays it was necessary to line up the social structure of an ancient civilisation, as filtered through early modern understanding, with the system designed for Shakespeare's time; such lining-up can never be more than approximate. This leads, for example, to the Triumvirs Caesar (i.e. Octavian), Mark Antony, and Lepidus being placed in category 0, and senators in category 2. Neither of these is quite precise historically; such decisions represent our best judgement given how Shakespeare presents these persons.

One final problem for the character metadata is assumed identities – that is, characters pretending to be someone they are not, whether as a theatrical performance, e.g. Bottom playing the role of Pyramus in *A Midsummer Night's*

Dream's wedding entertainment, or as a disguise, e.g. Portia disguising herself as Balthazar in *The Merchant of Venice*. There would be no clear justification to assign the words of Bottom-as-Pyramus to either Pyramus's category 1 or Bottom's category 5. Similarly, Portia-as-Balthazar's lines are neither clearly spoken by a man, nor clearly spoken by a woman. The solution to this involves the creation of a separate speaker ID for the utterances that are spoken under such an assumed identity, which therefore has a separate entry in the metadata that receives specially-flagged versions of the categories for the identity being performed. The flagged categories are assumed male, assumed female, and assumed problematic[14] for sex, plus an assumed version of each of the social categories (a0 to a7).

5 Word-level annotation

5.1 Spelling regularisation

A general consideration for corpus analysis is the way in which searches and frequency counts group together the different word forms in the corpus. For a lexicographical undertaking such as the *Encyclopedia* project, we normally wish to work with forms grouped according to their dictionary headword, so that an analysis of, say, verb *love* encompasses *love*, *loves*, *loving* and *loved* as well. This is accomplished using lemmatisation. But for historical texts, even the grouping of unlemmatised word forms is problematic, due to the extent of spelling variation. Of course, present-day English exhibits spelling variation to an extent; even ignoring UK versus US differences, we observe variation in, for instance, <z> versus <s> in the suffix *–ise*, or the use of open versus hyphenated versus closed (or 'solid') spelling of compounds (Sanchez-Stockhammer 2018:1–2). Working with modern corpora it is generally not detrimental to ignore this factor. But in Shakespeare's time, standardisation of English spelling was still decades at least away. In consequence, spelling variation is evident across most if not all words. This is highly problematic both for subsequent annotation, since both part-of-speech tagging and lemmatisation rely on lexicons with standard spelling, and for actual corpus analysis. For instance, a corpus query for *sweet* would miss *sweete*, one for *love* would miss *loue*, and one for *doubt* would miss *doute*. Likewise, the chances of these words being mistagged is much higher in their variant forms. Compounding the issue, some spellings are ambiguous (e.g. in early texts, orthographic *than* could represent either of *than* or *then*). Further, printer errors add to the variation (e.g. *aud* for *and*).

Our solution to this issue was to apply complete *spelling regularisation* (also called *standardisation* or *normalisation*, but the lack of standard spellings in this period makes these terms inappropriate) to the ESC corpora. Spelling regularisation is the process whereby any word detected as a non-standard spelling is replaced in the main text by its modern standard equivalent, the original variant form being retained in the markup. This can be accomplished either manually or automatically, although automatic approaches always involve some error rate (whether in false positives or false negatives). Given the complexity of the issue, and the critical importance of getting it right, we opted to regularise the texts of *ESC:Folio* manually. For this task we used the program *Variant Detector* (VARD2),[15] developed by various scholars at Lancaster University, most significantly by Alistair Baron (Baron and Rayson 2008). VARD2 automatically identifies variant forms by reference to a lexicon of standard forms. It affords two ways of handling the variants. First, in interactive mode, a user works manually through a text, determining how each variant form detected is to be treated using a variety of methods that exploit data acquired via the *training* that took place when running interactively. This allows the linguistic knowledge input into the system on one text to be utilised to automate work on subsequent texts.

The VARD2 interface, with *Henry V* loaded, is shown in Figure 1. In the primary text display, a yellow highlight indicates possible variants (words not matched in VARD2's lexicon of regular forms). When the user right-clicks on a highlighted word, automatically-generated suggestions for regularisation are presented, in order of probable correctness; most often the appropriate regularisation is on the list, and can simply be selected, but the user can also impose a change that is not among the suggestions. The remainder of the interface presents overview information and tools for examining the text's wordlist. When the text is saved to disk, the regularisations are represented in XML, as in the following example (from *King Lear*):

```
I must <normalised orig="loue" auto="false">love
</normalised> you, and sue to know you better.
```

XML-aware software can work directly with both regularised and unregularised spellings; software which simply ignores XML will work with the regularised spelling by default.

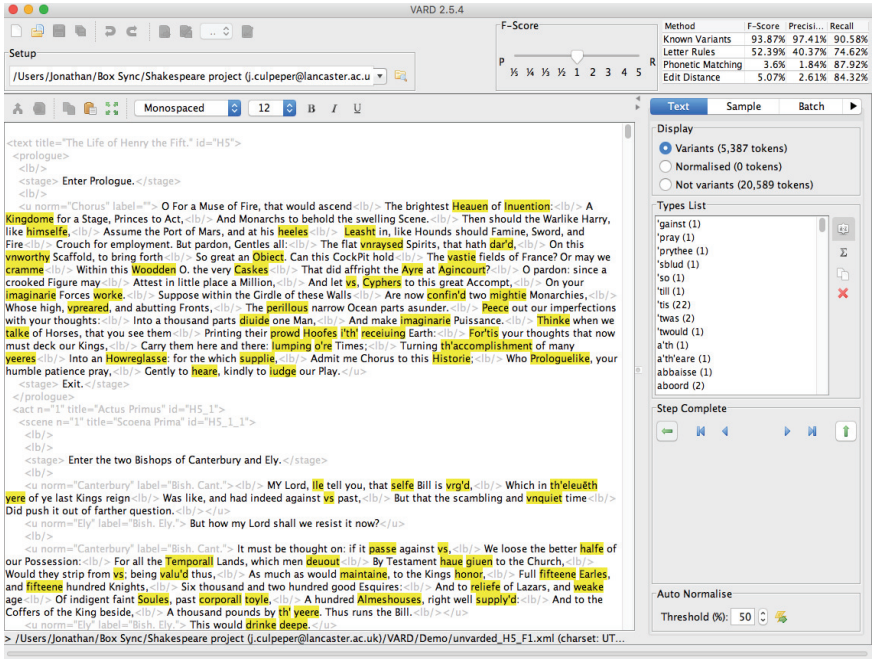

Figure 1: VARD2 in interactive mode, showing work-in-progress on Henry V

The advantages of the manual regularisation process (or 'Varding', as the project team quickly came to dub it) are many. First, it is possible to make regularisations that VARD2 either would not generate as possibilities, or would not score as sufficiently probable to implement automatically. Second, it is possible for real-word variants to be addressed (e.g. *dye* to *die*). Third, ambiguous forms can be resolved (e.g. *deere* to *deer* or *dear*). Fourth, issues that span multiple word tokens can be handled (e.g. *to morrow* to *tomorrow*).[16] Fifth, use of apostrophes can be made to conform to the modern pattern (e.g. *Majesties* to *Majesty's* or *Majesties'*). Sixth, outcomes of VARD2's default training model for automatic processing which for our needs are undesirable (e.g. regularising *sayth* and *saith* to *says*, rather than *sayth* to *saith*, our desired outcome) are not introduced to the corpus. Seventh, with VARD2 we can use the data from the manual regularisation of *ESC:Folio* as a new training model for automated mode – essential if the other ESC components were to be regularised in a manner compatible with *ESC:Folio*, since manual treatment of these other corpora

was not a workable prospect. In sum, all kinds of issues that would defeat automatic word-by-word processing can be handled.

However, there are also disadvantages. First is the inevitability of human error. Second, some regularisations (e.g. of place names) depend on world knowledge which any given analyst may or may not possess. Third, it can be difficult to ensure consistency between different analysts working on the texts, or even between the earlier and later work of a single analyst. We ameliorated this issue through in-advance discussion by the group undertaking the Varding to devise agreed guidelines for all to follow. However, this does not totally eliminate the problem; it will always be the case that some difficulties will emerge that this process did not anticipate. In consequence, extensive checking for consistency at later stages of analysis was needed. For instance, the treatment of *travail* and *travel* – forms that in F1 are both used for both of the modern words – was originally inconsistent. Some analysts were cautious about changing one to the other based on meaning; others were more confident. The whole set (88 instances in all) needed to be re-examined at a later stage so that all were regularised according to meaning where that was certain. Another example is the oaths *byrlady* ('by our lady') and *byrlakin* ('by our lady-kin'), both of which appear in multiple spellings, with and without apostrophe. During initial Varding, the spellings selected as the target of regularisation were *byr'lady* (with apostrophe) and *byrlakin* (no apostrophe). The inconsistency is easily accounted for: the decisions were made on different days and/or by different people, that is, without awareness of the other decision. Neither decision is on its face unreasonable or incorrect. But the inconsistency would be a lamentable flaw in the *Encyclopedia*'s lexicon or any other research output. This issue too was amended (by regularising to no-apostrophe forms) at the checking stage.

Having outlined the procedure, we should discuss the principles applied to the task of regularisation. While our approach was profoundly informed by that expounded by Archer et al. (2015), we did not hold back from departing from that system as and when the specific needs and goals of the *Encyclopedia* seemed to demand that we do so. Our primary principle was to respect the integrity of F1 (including Q1 for *Pericles* and *Two Noble Kinsmen*) as a historical object. So, for instance, there are a handful of instances of repeated words in F1 where the sense, and sometimes the metre, makes clear that that word should appear only once. Repeating a word, wholly or partially, in this way is known to be the kind of error a compositor could easily make (see Figure 2). Modern editions therefore typically drop the repeated word; we preserved it. More generally, while we did often consult modern editions (especially in the *Arden Shakespeare* series) when the appropriate regularisation of a form was unclear –

since to disregard the centuries of scholarship that underpin the footnotes of these editions would be foolish – we did not follow changes made by these editions that would make the text a less authentic representation of F1. Such changes often involve either hypotheses, however well-grounded, about lost forms underlying what is actually present in the text, or readings derived from Quartos preferred over F1 for one reason or another. An example of this is a line spoken by Mercutio in *Romeo and Juliet* which appears in F1 as "O that she were / An open, or thou a Poprin Peare". Modern editions typically amend *open* here either to "open Et cetera", which is what appears in Q1, or to "open-arse", which while not observed in Q1, Q2 or F1 is believed on good grounds[17] to be what was originally intended. We made neither change, retaining the F1 text – although the clear-cut spelling variants *poprin* and *peare* were both regularised, naturally. Again, post-Varding checks across frequency lists of regularised forms were needed to ensure that such editorial amendments were consistently eschewed.

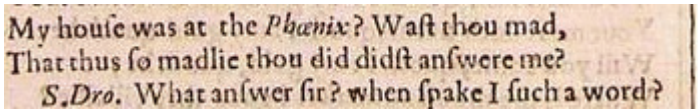
My houſe was at the *Phœnix*? Waſt thou mad,
That thus ſo madlie thou did didſt anſwere me?
S.Dro. What anſwer ſir? when ſpake I ſuch a word?

Figure 2: Compositor error in F1: did didst *for* didst *in* The Comedy of Errors

As with modern editions, at times we drew on information in the *Oxford English Dictionary*, e.g. to select a target regular form when more than one possibility exists; but again as with modern editions, we did not necessarily follow the OED where doing so would run against the principle of fidelity to F1. And in accordance with that principle, our ultimate fallback for ambiguous or indeterminate cases was to leave the word in question precisely as it appears in F1.

Our next major principle was that the regularisation process ought to preserve anything that was genuinely a feature of Early Modern English. So obsolete, archaic or rare words, e.g. *cozen/ed*, *haply*, and *morrow* were not replaced with non-obsolete synonyms; regularisation is not translation. However, in cases where such obsolete words exhibited spelling variation (which was usually the case) then a single form would be selected. Out of 35 tokens of *haply*, 8 originally had different spellings: *hap'ly*, *happely*, *happlie*, and *haplie*. Selection of the single target regular spelling, both for archaisms like *haply* and for other words with more than one candidate regularisation, was guided by four maxims:

- Regularise to one of the forms present in the raw text of the corpus, or, at least, a form that had currency in Shakespeare's time
- Avoid any regularisations that would be inconsistent with general regularisation practice across the corpus
- Regularise to the form that a contemporary English speaker would expect the word to take (so that words appear under the expected headword in the *Encyclopedia* and searching for the expected word finds the correct set of examples)
- Prioritise the candidate regular form which is most frequent in Shakespeare's work.

In the case of *haply*, these criteria are decisive: the spelling *haply* does occur in F1, it is the form that a user of the corpus or the *Encyclopedia* would most likely expect to be present (and is the spelling used as the OED headword), and it is the most frequent of the five different spellings. The form *haply* may be considered a *conventional archaism*: it is no longer really part of the productive vocabulary of today's English speakers except, perhaps, when 'putting on' an archaic tone, but its spelling is accepted by convention. English speakers today may not know to use *haply*, but many of them will be able to spot when it is spelt wrong!

However, cases where it is impossible to satisfy the criteria stated in all four maxims at once are not at all uncommon, and in these cases it was necessary to try to make a balanced judgement. Indeed, the hedge *wherever possible* might validly be appended to all four of the maxims as stated above. The most difficult cases were often dealt with via discussion across the team, true to the maxim of avoiding inconsistent practice; nevertheless, inconsistency did arise in different analysts' treatment of certain 'close call' situations, and had to be detected, and corrected, at the subsequent checking stage. There is, of course, no space here to discuss all the judgement calls that we made during Varding or during the subsequent consistency checks. Rather, we will consider some individual examples, and some classes of cases, where the principle of preserving genuine features of the Early Modern English vocabulary necessitated striking a balance across the four criteria.

One case which furnishes a sharp illustration of a conflict amongst maxims is *aye*. Contemporary English *aye* is normally an interjection, an archaism for (or dialect alternative to) *yes*. In Shakespeare, things are rather more complicated, as more than one word is involved.[18] As well as the interjection meaning *yes*, pronounced / əɪ /, there are cases in F1 of an adverb meaning *ever*, pronounced / ɛː /, and an interjection meaning *alas*, also pronounced / ɛː / (Crystal 2016: 38). Different spellings – mainly *aye*, *ay* and *I* – are used in a

somewhat but not entirely distinct manner: *I* is the usual F1 spelling for the *yes*-interjection (which was and remains homophonous with pronoun *I*); *aye* is the most common spelling for the two words pronounced /ɛː/; and *ay* is a rarer possibility for either, but more commonly represents the words pronounced /eː/. For *aye*-meaning-*yes*, following all the maxims listed above is simply not possible. The most frequent form is *I*, but this would absolutely not be expected by a modern speaker. Making the interjection homonymous with the pronoun would be confusing to corpus users expecting regularised spelling – and, not incidentally, just as confusing for automatic tagging software. That leaves *aye* and *ay* as target spellings. The former is never used for the *yes*-interjection in F1; the latter *is*, but (again using Crystal's numbers for the pronunciations) six out of eight cases of original spelling *ay* represent the *alas*-interjection. As for the factor of present-day expectations, our own intuition suggests the spelling *aye* for all of these – the distinction being partly or wholly eroded today, depending on the dialect. Less informally we note that the OED uses spelling *aye* for the *yes*-interjection, *ay* for the *alas*-interjection and presents both as alternatives for the *ever*-adverb; and that in the BNC 1994, there are about twenty times more examples of *aye* than *ay*, and swift inspection of the 251 cases of *ay* shows that many of them represent the *yes*-interjection. There is simply no obvious route through this tangle of factors. Respecting any imperative means breaching another. We opted ultimately to use the target spelling *aye* for all three of these words on the basis that the evidence on aggregate suggests that modern English speakers lack any consistent expectation of a separate word *ay* such that the expectation will be for just one spelling. Subsequent part-of-speech tagging and lemmatisation (see 5.2, 5.3) will cause the adverb to be a separate lemma from the two interjections. Meanwhile; the interjections will be grouped as one lemma – and, thus, a single dictionary headword – as is always the case for homophones with the same part-of-speech, such as noun *bank*. The *alas*-interjection and *yes*-interjection will subsequently be separated out as distinct usages in the *Encyclopedia*'s lexicon.

The extent of the discussion in the previous paragraph, compressed as it is, illustrates effectively how much effort was needed to make decisions on 'close call' words like *aye*. Even after all that discussion, we cannot claim that our decisions are definitive in terms of some Platonic ideal of correctly regularised Early Modern English. We can only claim them to be defensible, practically workable for purposes of the *Encyclopedia*, and as consistent with one another as could be humanly achieved.

An example of a 'close call' where even consistency had to be (reluctantly) put aside for the sake of other priorities is that of the pronoun/preposition func-

tions of *a*. As a preposition, this is a weak form for either *of* or *on*. As a pronoun, it is a dialectal weak form of *he*. Only a handful of plays (*Coriolanus* most prominently) use these weak forms at all frequently, and fourteen plays contain no instances of either; we may therefore suspect that their use over equivalent full forms is linked to compositor preferences. Both are usually spelt *a*; the preposition is also spelt *a'* when part of the contraction *a'th'*. The consistent thing to do here would have been to regularise all instances of both to *a*, the most frequent form as well as that preferred in the OED, making them homonymous with the indefinite article. Yet the need for consistent disambiguation seemed to us weightier than other principles here, because of the distinct treatment needed downstream at the lemmatisation stage (see 5.3). Generally, weak forms are lemmatised to corresponding full forms, so pronoun *a* needed to lemmatise to *he*. But preposition *a*'s full form is ambiguous (*on* or *of*); to avoid information loss we required it to be a lemma on its own. To make this possible, we imposed a somewhat artificial division of regular form at the Varding stage: *'a* for the pronoun, *a'* for the preposition, retaining *a* for the article. This is inconsistent with a number of our other practices (for instance, our usual preference for target regularisations not to include apostrophes: examples follow below) but in a way that we consider justified.

Two *classes* of word with no solution that fully abides by all our principles involve variation that persists into twenty-first century English. One is *-ise* versus *-ize*; both are present in F1, with the latter somewhat more common – and even extended to words where it is not etymologically motivated, e.g. *poize*, *practize*. Such unmotivated Z spellings obviously needed to be regularised to S. For the others, we discussed at length factors that should inform a policy for selecting the S or Z form as the target, such as the S-to-Z ratio for the particular form or lemma; but found that any policy designed around multiple such factors became too complex to be applied consistently in the context of manual Varding. In the end, as a practical matter, we decided that consistency was the critical factor for usability, and that thus a *simple* policy was needed. That policy: to regularise to S across the board. Users of the corpus (or *Encyclopedia*) might disagree with this call, but will, we believe, benefit in practical terms from our having made it.

The second such class is compounds spelt variably as two orthographic words ('open' compounds), or joined hyphenated, or joined unhyphenated ('closed' compounds). Generally speaking, where one of the latter is a possible regularisation target, it is preferable, because open compounds with spaces do not create separate types (or lemmas) in the corpus. For example, *to morrow* occurs 246 times and *tomorrow* only once in *ESC:Folio*, but we regularised all 247 to *tomorrow*, since if left unchanged, *to morrow* registers as an instance of

to and an instance of *morrow*, not as an instance of a single word. *Tomorrow* is straightforward enough: the modern form is clear, as is its status as an open compound in Shakespeare's time. But many other compounds were much less straightforward. In consequence, Varding led to inconsistencies in compound treatment across the corpus; to address them during checking, we queried the corpus to generate tables of all relevant cases and discussed them all individually to check that the decisions were neither arbitrary nor inconsistent. One particularly troubling issue was the question of whether certain bigrams were compounds at all. At least *one* hyphenated or fully closed example in F1 would be sufficient to accept as compounds other examples of the same thing but with a space. But often, even that was lacking.

Two typical cases were *mannor house* and *almes drinke*. Both are *hapax legomena* in F1. Both were initially identified as compounds, and Varded to *manor-house* and *alms-drink*, on the example of the *Arden Shakespeare*. On review, however, the justification for their compound status proved shaky. *Manor house* (thus spelt) is an OED headword, with examples of hyphenated *manor-house* from 1715. In *ESC:EEBO*, we find twenty earlier instances of hyphenated forms, with assorted other spelling variation; but we also find 127 open examples. Moreover, *manor house* is reasonably transparent and compositional: *manor* modifies (restricts) the head noun *house*. In light of this evidence, the case for *manor-house* as the appropriate regularisation in F1 seems much weaker than it would have to the manual Varder looking just at the running text of *Love's Labour's Lost*, with the *Arden* edition close to hand as a reference point. We reversed the original decision, regularising to *manor house*. *Alms-drink* proved to be a yet greater puzzle. The hyphenated spelling was supported by the OED as well as *Arden*; *alms-drink* is a sub-entry of *alms*. But there are just three examples: this one (in *Antony and Cleopatra*) and two hyphenated examples from the early 1900s, both apparently meta-linguistic: one is from a text about Shakespeare, the other an explicit discussion of the meaning of *alms-drink*! As the OED notes, this word's usage is "[c]hiefly after Shakespeare's use in [F1], which has been interpreted in various ways". A Google search is able to locate some other, non-reflexive uses, e.g. Rives (1888: 81), where *alms-drink* appears in a short story whose narration is intentionally archaic (and/or dialectal) throughout. But this 1888 literary example is not unlikely to have Shakespeare's use at its ultimate root (the story explicitly mentions Shakespeare twice). Without other early attestation of the hyphenated form, it would seem to be an artefact of relatively recent editors or scholars of Shakespeare. In other words, our initial justification for regularising to *alms-drink* stems from just this one example, which does not have a hyphen – and, like *manor house*, is transparent

as a bigram. We therefore changed the regularisation from *alms-drink* to *alms drink* – and extended the policy consistently to another several dozen instances of debatable compounds.

We applied the principles of regularisation defined above equally to matters of morphology. Variation representing actual features of Early Modern English grammar was preserved, but variant spellings of a single morphological form were regularised. So while for many weak verbs, the past tense or past participle suffix is rendered variously in F1 as *-d*, *-'d*, and *-ed*, this does not reflect a morphological difference, but merely different spellings of (free variants[19] of) the same morpheme; this was regularised per verb base.[20] By contrast, cases such as *holp/holpen* as the past tense/participle of *help* reflect a genuine diachronic change (of *help* from strong to weak inflection) and were not regularised to *helped* – even though *helped* does also occur in F1. But the spellings *holpe* and *holp* were merged via regularisation, because this distinction is not a matter of Early Modern English morphology. There is another maxim conflict here: although *holpe* is more common in F1 than *holp*, we targeted the latter, to abide by modern expectations and maintain consistency with our treatment of other words with supernumerary final letter E. A parallel case affecting nominal rather than verbal inflection is the irregular plural of *shoe*, which appears once as *shoone* and once as *shooen*; we regularised both to *shoon*, giving precedence to a modern 'conventionally archaic' spelling, preferred by the OED, over either actually-occurring form. A similar adjectival case is *horrider*, which was not regularised to *more horrid*.

We need not fully detail the logic of how the same approach led us to regularise, or not regularise, second/third person verb forms. Present *-(e)st* and *-(e)th* and past *-(e)d(e)st* – all sometimes but not always with apostrophes – were each mapped to a single spelling per verb base, favouring shorter non-apostrophe forms over longer forms wherever possible, e.g. *badst* over *bad'st* or *badest*. But third person *-(e)th* was *not* regularised to *-(e)s*.

Even on the morphological front, some difficult cases were not detected during Varding but only at the stage of consistency checking. For example, prefix *a-*, a derivative of the weak-form preposition discussed earlier, occurs in two main contexts: prefixed to present participles, reflecting their ancestry as preposition-governed gerunds (*a-wooing*, *a-hunting*, etc.); and prefixed to adjectives or nouns, usually forming adverbs (*a-tilt*, *a-nights*, *a-horseback*, etc.). Its two most frequent F1 spellings are as a separate word (e.g. *a hunting*) and prefixed with a hyphen; forms prefixed without hyphen (e.g. *abreeding*, *aworke*, *atilt*) are rarer. Since these words are not highly frequent either individually or collectively, their inconsistent treatment during manual Varding was not noticed. Dur-

ing the post-Varding check on hyphenated forms, the inconsistency was detected and, following discussion, resolved. Our regularisation target for all such words was the hyphenated form, even when that was less common than the version with a space. We thus respect the modern expectation that this prefix is not a separate word, but for consistency's sake, hyphenate even when a modern speaker might expect hyphenless prefixation. An additional motivation for this solution is that the OED seems to prefer hyphenated spellings as headwords when both are possibilities. We thus target *a-tilt* rather than *atilt*, for instance – like the OED. Yet even this policy had to be tempered by common sense for words such as *asleep* and *aright*, which are so firmly established as unhyphenated in today's English that it would be perverse to target *a-sleep* or *a-right* – especially since both are rendered without hyphen or space in the vast majority of instances in F1.

We wrap up our discussion of spelling regularisation by skimming over some more minor principles and strategies. Certain space-saving abbreviations of common words were expanded via the regularisation mechanism, including *frõ* to *from*, *yt* to *that* and *qd* to *quod*. Character name abbreviations appear only in stage directions (since the abbreviated labels at the start of speeches were moved into the markup) and where this occurred, they were likewise expanded, e.g. *Ros.* to *Rosencrantz*. One last technical principle we observed was to apply regularisation only to the text, not to any material that was moved into the markup. This included the *title* attribute of the <text>, <act> and <scene> elements (see 4.1), as well as, of course, the *label* attribute on <u>.

In some cases, we regularised forms plainly *intended* by Shakespeare to be irregular, a policy for which further justification is perhaps needed. In 4.1 above, we discussed words misspelt to indicate a dialect pronunciation. Very similar are deliberate corruptions and misuses, including malapropisms – irregular forms or spellings that indicate a character getting the word wrong. We dealt with these on a case-by-case basis. *Coram*, for example, was a common corruption of Latin *quorum*. It was thus mapped to the regular Latin form *quorum* (and wrapped in <foreign> tags). *Fartuous* is one of Mistress Quickly's famous malapropisms (*The Merry Wives of Windsor*, 2.2); it was mapped to *virtuous*, that is, the word she was attempting. Arguably, some information is lost in these changes. These misspellings are not mere irregularity, but meaningful acts of character representation. Yet on balance, and given that original forms are never deleted but merely moved into the XML, we thought it more important that searches and counts for *virtuous* should retrieve Mistress Quickly's attempt at the word.

Finally, non-English words were regularised according to the standard modern spellings (albeit without accents, to prevent any possible character encoding

incompatibilities). An example of this is the French-derived *adiew*; at the same time as being given <foreign> tags, it was regularised to *adieu*. For Latin, we targeted modern conventional spellings, as illustrated by the example of *Iouis* to *Jovis* presented in 4.1.

Perfect accuracy and consistency of any kind of manual annotation, spelling regularisation or otherwise, in a corpus of more than negligible extent is all but impossible to achieve, and we certainly cannot claim that *ESC:Folio* is without error in this respect. We do claim, however, that the procedure of manual regularisation, followed crucially by consistency checking on the basis of frequency lists of the regularisations performed, has produced a corpus with as few of these errors and inconsistencies as practically achievable.

5.2 Part-of-speech tagging

Part-of-speech (POS) tagging, the annotation of word tokens in running text with grammatical (strictly, morphosyntactic) category labels, is typically the first layer of annotation to be added to an English corpus,[21] or the second if the corpus has undergone spelling regularisation, as has *ESC:Folio*. That is, if a corpus has any word-level annotation *at all*, it will normally include POS tagging. This is for two main reasons: POS tagging is, along with lemmatisation, the easiest form of annotation to automate without introducing high error levels; and POS tagging is often a required input to other automated annotations including lemmatisation, semantic tagging, and constituency or dependency parsing. Since the *Encyclopedia* project absolutely required lemmatisation (see 5.1), we therefore required POS tagging.

The importance of POS tagging for lemmatisation in English is largely down to the polysemy of many lemmas across the major lexical word classes. That is, many words function as noun and verb (e.g. *love*) or noun and adjective (e.g. *Russian*) or adjective and verb (e.g. *elaborate*). The import of this hinges on the definition of *lemma* adopted, implicitly or explicitly, in corpus annotation, which is that a lemma is a set of *inflectionally* related forms. Since use of the same stem as noun/verb/adjective, or any pair thereof, is considered a derivational process (*zero-derivation* or *conversion*) and not an inflectional process, the two (or three) uses definitionally constitute two (or three) lemmas. For our purposes this implies two (or three) distinct entries in the *Encyclopedia*, to one of which every token of the word in question must be assigned for analysis, disambiguating the form for both POS and lemma. Thus, any token of *love* or *loves* tagged as a noun will be assigned to the lemma *love*-as-noun, and any token of *love*, *loves*, *loved* or *loving* tagged as a verb[22] will be assigned to lemma *love*-as-verb. We do not, however, wish to give the impression that this is merely a tech-

nical matter. It is also an issue of conceptual relevance for lexicographical work. Word meaning and usage develop in conjunction with grammar; the verb and noun *love* lemmas differ in more than just the structural slots they occupy in the syntax. Discourse functions, pragmatic weight, and even core lexical semantics occur in different patterns depending on a word's POS. Thus, it makes sense for the *Encyclopedia* to list POS-polysemous words as multiple entries, each entry analysing one lemma according to the set of tokens that exhibit the respective POS.

Early Modern English is distinct from present-day English in many respects, but at the level of grammar there are few enough differences that taggers developed for present-day English can successfully tag early modern texts reasonably well, once the issue of spelling variation has been addressed – albeit with some degradation of accuracy (Rayson et al. 2007). (This is in sharp contrast to both Middle English and Old English, of course.) POS tagging is a long-established technology; at Lancaster University, the UCREL research centre's in-house tagger for English has been under development for over forty years. This tagger is the *Constituent Likelihood Automatic Word-tagging System*, better known as CLAWS[23] (Garside et al. 1987; Leech et al. 1994). Different POS taggers utilise different systems of grammatical category labelling (*tagsets*). In the tagsets used by CLAWS, every category is represented by an alphanumerical code, such as NN1 for singular common noun, or VV0 for base-form lexical verb. (Henceforth, we present POS-tagged words using the longstanding convention *word_TAG*, e.g. *love_NN1* or *love_VV0*, although in point of fact our actual implementation does not use this format.) We use the CLAWS6 (or C6) tagset.[24] Both the design of POS tagsets, and the implementation of POS tagging software, are fields unto themselves to which we cannot hope to do justice here. Briefly, however, CLAWS works as follows (see Garside and Smith 1997:111ff for a full account). First, CLAWS analyses each word token by looking it up in a lexicon that associates word forms with their possible POS tags (e.g. *love* with NN1 and VV0), and assigns all possible tags to the token. If the form is not in the lexicon, guesswork is applied; for instance an unknown word ending in *<ed>* is likely to be a past tense/past participle form with possible tags VVN, VVD and JJ.[25] Second, CLAWS picks one of the assigned tags as the most likely to be correct given the context in which the token occurs. This judgement is based on a method called a *(hidden) Markov model*, which estimates probabilities for the different tags based on a matrix of sequencing probabilities (e.g. the likelihood that the word following an adjective will be a noun). In some cases, a word's tags are pre-marked as more or less likely in the lexicon, and the Markov model also incorporates this information. Finally, CLAWS corrects the outcome of the

probabilistic disambiguation in cases where a sequence of tokens is recognised as an instance of a known type of multi-word unit or phrase pattern. These sequence types, referred to collectively as *idioms*, are listed together with their tags in a separate lexicon.

CLAWS typically achieves 96–97 per cent accuracy on written texts, being slightly less accurate on spoken texts (Garside and Smith 1997:118). It is the need to match word forms in the lexicon that necessitates prior spelling regularisation. Without decent regularisation, running CLAWS on texts in Early Modern English produces very poor results indeed. Early work using VARD as an input to CLAWS, conducted by members of the *Encyclopedia* project team, showed that, even with regularised Shakespeare texts, CLAWS only achieved 89 per cent accuracy (Rayson et al. 2007). While a drop from 96 to 89 per cent might not seem much, 89 per cent accuracy is roughly one error every ten tokens. Worse, these errors are not spread evenly, but are concentrated in the content words (since function words are usually more straightforward) and *especially* content words unknown to the tagger, normally lower-frequency items. But lower-frequency items represent the overwhelming majority of the entries in a lexicon. 89 per cent accuracy, though not a bad place to start from, is simply not good enough for the *Encyclopedia*. We thus engaged in substantial effort to improve the accuracy of the POS tagging.[26]

Using manual rather than automated regularisation with VARD2, as previously explained, was a major component of our strategy for better POS tagging. In corpus annotation, *errors cascade downstream*, so that any token regularised incorrectly – or incorrectly left *un*regularised – is likely to be tagged and lemmatised incorrectly too. Like spelling regularisation, it is possible for POS tagging to be done manually, but this process is far too labour-intensive for us to even consider it. An alternative is *post-editing*: tasking the output from CLAWS and manually examining it and hand-correcting errors. This is also highly time-consuming: to correct one million words, the rough size of *ESC:Folio*, is about at the limit of what is feasible within the span of a project like the *Encyclopedia*. We determined, therefore, that while post-editing *ESC:Folio* was necessary, we would adjust the operation of CLAWS to minimise the requirements for manual fixes. The strategy we adopted was to make changes and additions to the CLAWS lexicon and idioms, and marginally to the tagset, but to leave unaltered the probability matrix developed for contemporary written English and originally trained on the British National Corpus 1994 (Leech et al. 1994). The kinds of grammatical cues that this matrix captures (e.g. words after articles tend to be adjectives or nouns; nominative pronouns tend to be followed by finite verbs; degree adverbs precede adjectives) are among the features of English which

have changed *least* over the past four hundred years. We also wrapped the operation of CLAWS in pre-processing and post-processing stages which addressed specific linguistic features of the texts or in some cases particular requirements of the *Encyclopedia*.

By contrast to the basic grammar of the language, the aspect of English that *has* changed noticeably since Shakespeare's time – spelling aside – is lexis: individual words (and phrases) and their particular usage and grammatical behaviour. Adjustments to the CLAWS lexicon addressed many of these issues. Let us consider a few representative examples. Some words that have disappeared from English over the last four hundred years, e.g. *wot* 'know' and *iwis* 'certainly', had to be added to the lexicon. Our modified CLAWS lexicon covers around 3,300 additional word forms in all. Other words have changed in grammatical behaviour; for instance, *fee* is only a noun in contemporary English but could equally well be a verb in Early Modern English. In yet other cases a word's possible tags needed to be assigned different probability weightings. For example, *smart_VV0* and *smart_NN1* are flagged in the CLAWS lexicon as drastically less probable than *smart_JJ*, but this proved inaccurate for Early Modern English; e.g. in *ESC:Folio*, there are exactly three of each. The lexicon entries for *fee* and *smart* were thus amended to add *fee_VV0* and to weight evenly each possible tag for *smart*. The probability profiles of approximately 80 common words were changed in this way.

On another front, we enhanced CLAWS' ability to separate out pronoun-verb contractions – or, more formally, to introduce token boundaries between clitics and their host words. For purposes of POS tagging, it is usually preferable to separate out clitic forms so that they can receive a separate tag to their host. This allows further annotation processes to treat clitics identically to corresponding full forms. For instance, if enclitic *'re* is not separated out of *they're*, it cannot receive the same tag as *are* (which is VBR); nor can it be lemmatised as *be*. Combinations of pronoun plus enclitic verb that exist today are accurately handled by CLAWS already (e.g. *he's*, *she'll*, *I'm*, *you've*); of course many of these are also present in F1. But there were more such combinations in Shakespeare's day (e.g. *thou'dst*), including some where the pronoun is cliticised onto the verb rather than vice versa (e.g. *methinks*, *quotha*) and for the sake of consistency CLAWS tokenisation rules were added for these contractions.

CLAWS using a tagset developed for twentieth-century English causes its output to overlook assorted grammatical features of Early Modern English. For instance, as marginal phenomena in contemporary English, *thou* and *thee* are treated by CLAWS merely as alternatives to *you*. But in Shakespeare's time, the distinction was both morphosyntactically significant, because of the distinct verb agreement triggered by *thou* as opposed to *you*, and pragmatically signifi-

cant, in terms of the local construction of social status and/or closeness (see, *inter alia*, Busse 2002). We therefore extended the tagset, creating an Early Modern English variant on the C6 schema, with distinct tags for these pronouns: *thou_PPYS1* and *thee_PPYO1*. Similarly, in CLAWS by default second person singular verb forms are assigned the same tags as the equivalent uninflected form: present-tense *want_VV0* and *wants_VVZ* alongside *wantest_VV0*. We added new tags, distinguished by final T as a mnemonic for the final letter of suffix *-(e)st*, for these inflections. The result was tagging such as *wantest_VVT* (second person singular, present tense) and *wantedst_VVDT* (second person singular, past tense) for lexical verbs, plus parallel tags for primary and modal auxiliaries, e.g. *art_VBT*, *hadst_VHDT*, *canst_VMT* instead of VB0, VHD, VM. These novel tags were overlaid onto the CLAWS output *after* completion of normal processing. This meant that affected words had their default tags (i.e. *art_VB0*, *hadst_VHD*, *canst_VM*, etc.) at the point of probabilistic disambiguation, the matrix naturally lacking entries for our new tags. In consequence, though many of the extra 3,300 entries are second person singular verbs, in the lexicon these forms are associated with the default CLAWS tags, not our extended tags.

The POS-tagged *ESC:Folio* texts were then *manually post-edited*. This is exactly what it sounds like: the entirety of each text is inspected by a human trained in use of C6, checking each and every tag for correctness, and amending all errors. Figure 3 shows what this looks like in practice. A fast post-editor can check through a Shakespeare play in about the time it would take to read that same play carefully. This considerable investment of time was clearly worthwhile for the core dataset, but was not feasible for our other corpora (*ESC:EEBO*, the largest, is three hundred million words). However, those corpora did benefit from further improvements to CLAWS made in light of the post-editing work.

The main risk of post-editing is human error, normally omission of needed corrections rather than insertion of new errors, though the latter does happen. A secondary risk is inconsistency. Unlike CLAWS, which given the same input always produces the same output, a human being faced with a dilemma over a tag might choose one option on one occasion, but the other on another occasion. The best protection against these risks is for two post-editors to independently process all texts, and then debate and resolve any differences; however, this doubles the labour required. With limited time available, we used alternative approaches to mitigating human error.

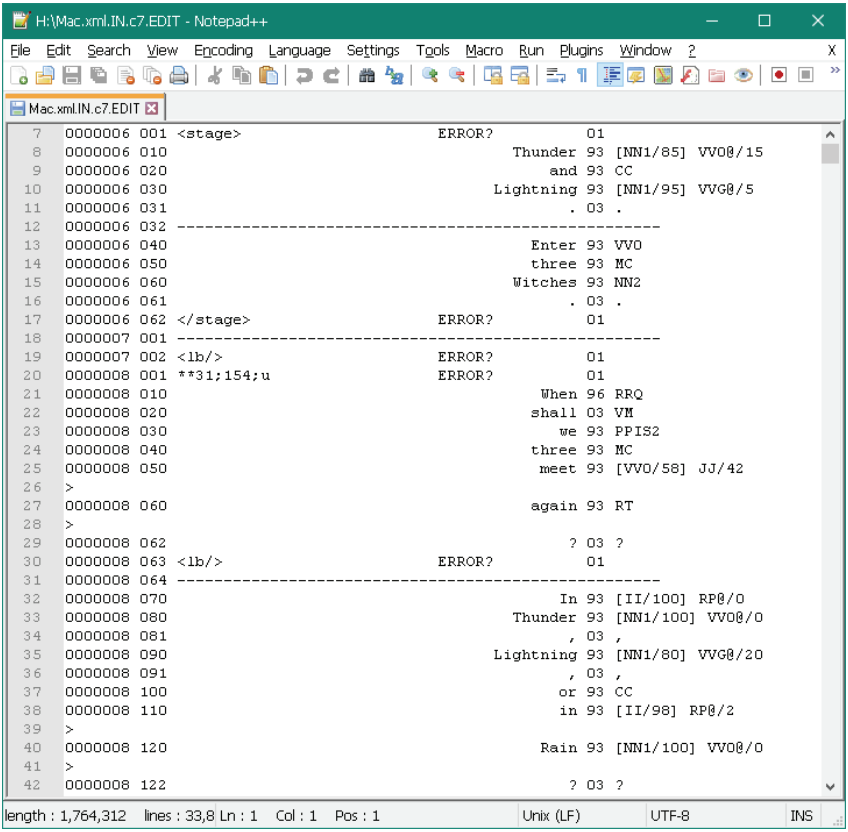

Figure 3: Post-editing of POS tagger output. This screenshot shows the beginning of Macbeth *in the format used for post-editing (raw CLAWS output with minor readability tweaks), open in Notepad++. All but one visible token is correctly tagged. The exception is* meet*, which should be tagged VVI (infinitive) not VV0 (present tense).*

First, we made the decision to limit the errors to only the major word class level of the C6 tagset. That is, a word tagged VV0 (verb) that was actually NN1 (noun) would be corrected; but a word tagged VV0 (present tense) that was actually VVD (past tense) would not. By limiting the classes of errors to be addressed, we lowered the risk of inconsistency, and freed up time for other checking techniques. Admittedly, for some forms this had to be revisited at the

lemmatisation stage. For instance, *wilt* can be either VMT (second person singular of *will*) or VV0 (base form *wilt*, as in, a plant wilting), both verb tags; to correctly assign lemmas *will* and *wilt* respectively, the second-level VM-versus-VV distinction is needed. But such forms are the minority.

Second, we decided that each text would pass through three post-editors, but that rather than repeat each other's work, they would apply different techniques in serial. The first post-editor worked with a list of known, common tagger errors (e.g. complementiser *that* versus determiner *that*), and searched each text for tokens liable to those errors – but did not read the whole text. The second post-editor performed a pass of the text in less than full detail, with the goal of fixing as many other straightforward or obvious errors as possible. The workload of the third post-editor, who performed a full in-depth pass, was thus reduced as much as possible, letting them focus on the most difficult cases of error, plus where possible fixing mistakes of omission or commission made by the first two analysts. This procedure had the further advantage that only the final post-editor needed a fully detailed knowledge of C6; the first two were able simply to pass over any issues that were beyond a sound but basic level of expertise with the tagset.

Post-editing does not completely guarantee accuracy of tagging, since human error cannot be wholly eliminated. Even putting errors aside, there do occur constructions on which different human grammarians might disagree even *after* discussion (although this is true for modern texts as well). But we can state with confidence that the grammatical tagging in *ESC:Folio* is as close to 100 per cent accuracy as humanly possible.[27]

5.3 Lemmatisation and semantic tagging

The importance of the lemmatisation process to the *Encyclopedia* project has already been discussed at length. It suffices here to briefly recount some practical aspects.

Text tagged by CLAWS is normally then passed to a second tagger, USAS,[28] the *UCREL Semantic Annotation System* (Wilson and Thomas 1997; Rayson et al. 2004). As its name indicates, USAS is primarily a semantic tagger, but it also performs lemmatisation – since USAS analyses lemmas rather than word forms, this is a necessary design. These lemmas are linked with top-level POS categories to distinguish, say, *convert_NOUN* from *convert_VERB*. For *ESC:Folio*, we ran the text through USAS after post-editing of the CLAWS output was complete. For *ESC:Verse* and *ESC:Quartos* (plus *ESC:Comp* and *ESC:EEBO*), the output of our modified CLAWS system was passed straight to USAS. As with CLAWS, while making no changes to USAS itself, we made modifications to

the lemmatisation resources (a lexicon mapping word forms plus POS tags to lemmas, and a set of rules for guessing the lemma of words not in the lexicon) on the basis of the requirements of the *Encyclopedia* project and the language of the period, for instance to make sure that *mayst_VMT* would be lemmatised as *may_VERB*. The output from USAS was passed through a final post-process to generate two end-state versions of each corpus: the columnar format required by CQPweb, and the XML format for corpus distribution.

This ultimate version retained the semantic tags produced by USAS. However, it is worth noting that we did not make any effort to improve, or adjust to the period, the semantic tagging as we did the POS tagging and lemmatisation (although of course, since POS tagging and lemmatisation are both upstream of semantic tagging, our efforts there will have had some impact on the semantic tagging). This was because substantial work to enhance semantic tagging for the early modern period has already been undertaken by the SAMUELS project,[29] which developed the *Historical Thesaurus Semantic Tagger* (HTST) (Alexander et al. 2015; Piao et al. 2017). The *Encyclopedia* project had no reason to duplicate this effort. On the contrary, applying the HTST to the ESC datasets is an undertaking which we hope to explore in future work.

6 Conclusion

One might say, somewhat provocatively, that the research reported in this paper has accomplished nothing *new*. Many previous corpora have been enhanced in terms of the texts that comprise them, the way they are organised for corpus analysis, and the annotation that is applied. In fact, Lancaster has a long tradition in corpus enhancement, one initiated in particular by Geoffrey Leech's work on the annotation of the (original) British National Corpus (see 5.2). However, historical texts and the language of Shakespeare present an extreme – a concentration of challenges on multiple dimensions that demand innovation.

Ours is certainly not the first corpus to represent the works of a single writer, but Shakespeare's works constitute a particular challenge even at the very initial stage of text selection, as there is no definitive list of what constitutes his works (as discussed in Section 2). We have explained how the specific needs of our research agenda – the need for a single, stable corpus for analysis; the need for an actual historical textual entity as an anchor point; the need for a source text free from editorial interference – determined the selection of content for *ESC:Folio*, and more briefly, for two minor components of the ESC. As we have detailed, our base texts were derived from the transcriptions created by ISE. These files already had XML markup; we have explored (in 4.1) the ways in

which we reworked the XML structure to align it with the requirements of corpus analysis. Use of XML in this way is definitely not new, but novel challenges of implementation have been addressed in the conversion between XML schemata with very different purposes and very different structures and element vocabularies. Using speaker metadata to enrich the analytic affordances of a corpus is likewise not new – and the technical details of how we did it in fact differ very little from those of the first major corpus to accomplish this using XML,[30] that is, the spoken part of the British National Corpus. The challenge here lay in the definition of a metadata scheme for characters in early modern drama able to sense of the typical demographic criteria of sex and class in the context not only of a four-centuries distant social setting, but also of fictional worlds as depicted from that setting: worlds where, among other things, persons assume identities that can shift both sex and class for disguise or for performance. We believe that the advancements we have made upon Archer and Culpeper's (2003) earlier work on this form of sociopragmatic metadata, at least with respect to speaker values, constitute a contribution to this field.

The layers of analytic annotation that we have applied to *ESC:Folio* and the other corpora – spelling regularisation, part-of-speech tagging, lemmatisation and semantic tagging – are by no means new. Of course, English spelling is not fully regular even today, and in some genres (e.g. some forms of electronic media) it is distinctly *irregular*. Nevertheless, historical texts from Shakespeare's era, with spellings from a good fifty years before standardisation really took hold and further complicated by the vagaries, idiosyncrasies and errors of multiple compositors, take things to a whole other level – as the account we have given here (in 5.1) has, we hope, made abundantly clear. Earlier work to apply automated taggers for contemporary English to historical texts by means of pre-annotating spelling variation using VARD2 or similar software has had some success. But both the nature of the Shakespeare plays, and the needs of the *Encyclopedia* project as a whole, raised the impact of the challenges with annotation to a drastically higher level, requiring new solutions above and beyond those implemented in the prior literature.

Historical lexicography requires a *highly* accurate headword (lemma) list, and a *highly* accurate accounting of what forms are linked to each lemma, and a *highly* accurate mapping to each form of the variant spellings that represent it in the raw text. The circumstances of Shakespeare's time sharpen these factors yet further. Though the lexicon of any language is always evolving, Shakespeare's time was an extreme. It was in this period that borrowing from other languages (especially Latin, but also French, Italian, and so on) was at its peak. In addition, internal word formation was in a state of flux, as we noted in our discussion of

compounds. The English of Shakespeare's day was in a state of grammatical transition, with morphological forms and syntactic structures that were archaeological relics of earlier stages of English, yet also with new forms and structures. These grammatical matters cannot be set aside from work on the lexis, because they are directly relevant to POS tagging and thus to the aforementioned complex of issues around lemmatisation and thus lexicography. The solutions to all these challenges that we have presented here, in going beyond the (successful) strategies deployed in prior work, therefore constitute a distinct contribution to the field. If the problems we have addressed are not novel, then the degree to which we needed these issues to be ameliorated is. And, as we have shown repeatedly in this paper, it *is* possible to surmount these known challenges at the extreme level they arise for the plays of William Shakespeare. The solutions we have discussed will, we hope, be of use to other researchers engaged in historical linguistic endeavours with corpora where the same challenges rise to that same extreme.

We conclude this paper with some thoughts about possible research exploiting the *ESC*. These thoughts are presented as an overlapping, open-ended list clustered under four broad headings: Shakespeare's language and style, lexicography, grammar and social variation. Within each of the clusters the spread of topics is influenced by the fact that the *ESC* enables contributions of three kinds: to the study of Shakespeare (*ESC:Folio*), the study of early modern playwrights (*ESC:Comp*), and the study of Early Modern English (*ESC:EEBO*).

- *Shakespeare's style and its distinctiveness*: This cluster includes, without being limited to, the research agenda of authorship attributions scholars. This may focus, for example, on whether Shakespeare wrote some given text *X* as opposed to some other text *Y*. Yet attribution scholars have begun to move beyond that narrow focus and to explore broader aspects of style (see Craig and Greatley-Hirsch 2017). Potential research questions that the *ESC* can help address include: (1) How did Shakespeare use the various linguistic components of style, in terms of words, phrases, themes and other features? (2) How does this linguistic usage vary, across plays, genres, different kinds of characters, and other dimensions? (3) How distinctive is Shakespeare's linguistic usage in comparison with that of contemporary playwrights? (4) How does his linguistic usage compare with that of contemporary writers of all kinds?
- *Lexicography*: Dictionaries of Shakespeare exist – Onions ([1911] 1986) and Crystal and Crystal (2002), to mention only two. However, none of these are founded on corpus methods, the methods of choice for today's

dictionary makers.[31] A fully corpus-based dictionary has yet to be realised. Of course, there is no reason to stop there – why not a dictionary of early modern playwrights, or a dictionary of Early Modern English? More radically, we can envisage a hybrid of these possibilities, such as a dictionary focusing on Shakespeare, but incorporating comparisons with the lexical usage of early modern playwrights (and/or writers in general).

- *Grammar*: The current state of published grammars of Shakespeare's language parallels the current state for dictionaries, in that some respected but very old work has been fairly recently superseded by modern undertakings that do not, however, lean heavily on corpus data (as do many major grammars of today's English). Abbott (1870) is the original Shakespeare grammar; Hope's (2003) grammar is an explicit successor to, and sometimes critique of, the work of Abbott. Somewhat more substantial and nuanced than either of those is Blake (2002). However, none deployed the full arsenal of corpus methods. The grammatical tagging of the *ESC* is a critical utility enabling use of these methods in the service of detailed research on grammar. As with a dictionary, a full grammar could focus just on Shakespeare's language, on playwrights' language, on all writing of the period, or on some hybrid of these.
- *Social variation*: Although literary criticism abounds with comments on Shakespeare's characters, the issue of how social variables such as sex or status impact on the language put into those characters' mouths has not yet been comprehensively and systematically studied, and certainly not via corpus methods. Our metadata classifying 'speakers' according to these social factors, applied to both *ESC:Folio* and *ESC:Comp*, will enable considerable advances On this front. Interestingly, the question of language and social variation *generally* in Early Modern English is relatively well served, notably by scholars linked to the University of Helsinki (e.g. Nevalainen and Raumolin-Brunberg 2003). However, that work has been largely based on one genre, that of letters, as represented by the *Corpus of Early English Correspondence* compiled by the aforementioned team at Helsinki. The results which we anticipate emerging from study of drama through the *ESC* will therefore provide a fruitful complement to that work.

As might be imagined, the *Encyclopedia of Shakespeare's Language* Project team is already pursuing many of the above research possibilities. The volumes of the *Encyclopedia* will begin to be published from 2022; academic dissemination has begun with papers published in a special issue of *Language and Litera-*

ture 29(3). But, needless to say, a corpus such as the *ESC* opens the door to far more potential research than could possibly be undertaken by one team, and we hope other researchers will soon begin to fully avail themselves of its riches.

Acknowledgement

The research presented in this article was supported by the UK Arts and Humanities Research Council (AHRC), grant reference AH/N002415/1.

Notes

1. For information: https://wp.lancs.ac.uk/shakespearelang/. For access: http://corpora.lancs.ac.uk/esc-user-service/.
2. F1, Q2, etc. are standard scholarly abbreviations for different extant editions of plays by Shakespeare: 'First Folio', 'Second Quarto', and so on *mutatis mutandis*. The Folios are large 'collected editions' whereas the Quartos (and, occasionally, Octavos) are individual plays published separately.
3. https://www.whatsonstage.com/london-theatre/news/rsc-gregory-doran-shakespeare-authorship_50845.html (last accessed 16th March 2021).
4. For the 18 plays with a Quarto edition predating a 1623 appearance in F1, it is the date of the former, rather than of the F1 text used in the corpus, that is given as the main publication date; the publication date of the specific source edition is given in an additional metadata field.
5. https://internetshakespeare.uvic.ca
6. https://internetshakespeare.uvic.ca/Foyer/quality.html (last accessed 24th February 2021).
7. https://internetshakespeare.uvic.ca/Foyer/plays/ (last accessed 24th February 2021).
8. *All's Well that Ends Well, The Comedy of Errors, Measure for Measure, The Merchant of Venice, Two Gentlemen of Verona, and Twelfth Night*.
9. A reviewer of this paper wonders whether the Oxford English Dictionary might assist in establishing a word's 'foreign' status. We did check the OED, but in most cases it did not give the fine-grained detail required.
10. We used the term *speaker* in accordance with general methodologies for spoken corpora. Nevertheless at no point do we forget that, in truth, we are dealing with characters in a fictional world. Similarly, our annotation of speech within *utterance* tags as discussed previously does not reflect any disregard of the fact that, in truth, these are *speeches* written by Shakespeare to be spoken by actors.

11. This implies that any character referred to as, for instance, a *Messenger*, receives the same ID throughout any given play – even though there may be no reason to imagine that Shakespeare actually intended such a Messenger to be the same person across different scenes (nor even that he gave much thought to the matter at all). Grouping same-name incidental characters in this way ensures their utterances receive the correct social categorisations, but does lose some information. Future work to be published within the *Encyclopedia* itself on social networks in the plays will distinguish these characters.
12. For instance, the text of *Macbeth* makes clear that Duncan is elderly, and Malcolm and Donalbain are young adults. But for Macbeth, Lady Macbeth, Banquo, and Macduff the most that can be said is that they are somewhere in the range between those age groups. For other characters (Lennox, Ross, the Porter) not even that much is apparent.
13. See, for example, Holmes (1982), Wrightson (1982, 1991), Sharpe (1987), Corfield (1995) and Hunt (1996).
14. 'Assumed problematic' sex only applies to two non-human assumed identities in *A Midsummer Night's Dream*: Snout performing the Wall, and Snug performing the Lion.
15. http://ucrel.lancs.ac.uk/vard/.
16. VARD2 supports regularisation by joining words together only in manual mode (in automatic mode, permitting joins would make the task of assessing all possible regularisations exponentially more difficult). A minor technical point is that default XML output from VARD2 uses a <join> tag to indicate such manually merged forms, but for our purposes, it was sufficient to use the <normalised> tag and its *orig* attribute (see 4.1) for joins as well as all other regularisations.
17. First, a one-syllable word is needed here for the metre. Second, *open-arse* is an archaic/dialectal name for the medlar fruit, which is established as a topic of discourse earlier in the same speech. Third, Mercutio's dirty joke only works if the word is *open-arse*.
18. Indeed, our discussion here does not account for all the complexities of *aye* and related words.
19. By 'free variants' we mean the syllabic and non-syllabic forms of this suffix whose interchange is driven by metre rather than allomorphy.
20. This discussion elides for brevity the issue of *-t* as a spelling of the [t] allomorph in verbs for which this is not an accepted spelling today: *blest, likt* for *blessed, liked*.

21. A minor caveat: conceptually, tokenisation – the splitting of input data into units for analysis (tokens) – precedes POS tagging, but in practice tokenisation is almost always performed as part of the POS tagging task.
22. Yet further lemmas are created when *loved* and *loving* are used as adjectives, of course.
23. http://ucrel.lancs.ac.uk/claws/.
24. http://ucrel.lancs.ac.uk/claws6tags.html.
25. VVN: past participle form of lexical verb. VVD: past tense form of lexical verb. JJ: adjective, in this case, an adjective zero-derived from the verbal participle.
26. The work of modifying the CLAWS resources and post-editing the results was led and mostly undertaken by Andrew Hardie, with input from Jane Demmen.
27. We plan a paper dealing in more detail with the difficulties of tagging early modern Shakespearean data.
28. http://ucrel.lancs.ac.uk/usas/.
29. http://www.gla.ac.uk/samuels/.
30. A technical note: the British National Corpus 1994 predates XML. Its first two editions were in SGML, the markup language ancestral to XML, rather than XML itself. This distinction is not relevant to the point at hand.
31. We understand that Crystal and Crystal (2002) did use some concordances.

References

Abbott, Edwin, A. 1870. *A Shakespearian grammar.* Third edition. London: Macmillan.

Alexander, Marc, Fraser Dallachy, Scott Piao, Alistair Baron and Paul Rayson. 2015. Metaphor, popular science and semantic tagging: Distant reading with the Historical Thesaurus of English. *Digital Scholarship in the Humanities* 30(suppl_1): i16–i27. https://doi.org/10.1093/llc/fqv045

Archer, Dawn and Jonathan Culpeper. 2003. Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics. In A. Wilson, P. Rayson and A.M. McEnery (eds.). *Corpus linguistics by the lune: A festschrift for Geoffrey Leech*, 37–58. Frankfurt/Main: Peter Lang.

Archer, Dawn, Merja Kytö, Alistair Baron and Paul Rayson. 2015. Guidelines for normalising early modern English corpora: Decisions and justifications. *ICAME Journal* 39: 5–24. https://doi.org/10.1515/icame-2015-0001

Baron, Alistair and Paul Rayson. 2008. VARD 2: A tool for dealing with the spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, U.K.*, 22 May 2008.

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

Bray, Tim, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler and François Yergeau (eds.). 2008. *Extensible Markup Language (XML) 1.0.* Fifth edition. W3C Recommendation 26 November 2008. https://www.w3.org/XML/

Busse, Ulrich. 2002. *Linguistic variation in the Shakespeare corpus: Morpho-syntactic variability of second person pronouns*. Amsterdam: John Benjamins.

Corfield, Penelope J. 1995. *Power and the professions in Britain, 1700–1850*. London: Routledge.

Craig, Hugh and Brett Greatley-Hirsch. 2017. *Style, computers, and Early Modern drama: Beyond authorship*. Cambridge: Cambridge University Press.

Crystal, David. 2016. *The Oxford dictionary of original Shakespearean pronunciation*. Oxford: Oxford University Press.

Crystal, David and Ben Crystal. 2002. *Shakespeare's words: A glossary and language companion*. London: Penguin.

Demmen, Jane. 2020. Issues and challenges in compiling a corpus of early modern English plays for comparison with those of William Shakespeare. *ICAME Journal* 44: 37–68. https://doi.org/10.2478/icame-2020-0002

Duncan-Jones, Katherine and H.R. Woudhuysen (eds.). 2007. *The narrative and other poems* (The Arden Shakespeare) [also titled: *Shakespeare's poems: Venus and Adonis, The Rape of Lucrece and the shorter poems* and *Shakespeare's poems*]. London: Bloomsbury.

Farmer, Alan B. and Zachary Lesser (eds.). 2007. *DEEP: Database of Early English Playbooks*. Available online at http://deep.sas.upenn.edu. Last accessed 24 February 2021.

Garside, Roger and Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech and T. McEnery (eds.). *Corpus annotation*, 102–121. London: Routledge.

Garside, Roger, Geoffrey Leech and Geoffrey Sampson (eds.). 1987. *The computational analysis of English*. London: Longman.

Gledhill, Christopher, J. 2000. *Collocations in science writing*. Tübingen: Narr.

Hardie, Andrew. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal* 38: 73–103. https://doi.org/10.2478/icame-2014-0004

Hinman, Charlton (ed.). 1968. *The Norton facsimile: The First Folio of Shakespeare*. New York: Norton.

Holmes, Geoffrey S. 1982. *Augustan England: Professions, state and society, 1680–1730*. London: George Allen and Unwin.

Hope, Jonathan. 2003. *Shakespeare's grammar*. London: Arden Shakespeare.

Hunt, Margaret R. 1996. *The middling sort: Commerce, gender, and the family in England, 1680–1780*. Berkeley: University of California Press.

Kelly, Erin (ed.). (n.d.). The Taming of a Shrew. Internet Shakespeare Editions. Available online at https://internetshakespeare.uvic.ca/Library/SLT/plays/the%20taming%20of%20the%20shrew/ashrew.html. Last accessed 11 March 2021.

Leech, Geoffrey, Roger Garside and Michael Bryant. 1994. CLAWS 4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, Kyoto, Japan, 622–628. http://ucrel.lancs.ac.uk/papers/coling1994paper.pdf

McEnery, Tony and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Murphy, Sean. 2019. Shakespeare and his contemporaries: Designing a genre classification scheme for Early English Books Online 1560–1640. *ICAME Journal* 43: 59–82. https://doi.org/10.2478/icame-2019-0003

Nevalainen, Terrtu and Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics*. London: Longman.

Onions, Charles T. 1986 [1911]. *A Shakespeare glossary*. Second edition. (Enlarged and revised by Robert D. Eagleson). Oxford: Clarendon Press.

Piao, Scott, Fraser Dallachy, Alistair Baron, Jane Demmen, Steve Wattam, Philip Durkin, James McCracken, Paul Rayson and Marc Alexander. 2017. A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech and Language* 46: 113–135. https://doi.org/10.1016/j.csl.2017.04.010

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nick Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics 2007, July 27–30, University of Birmingham, UK*. http://ucrel.lancs.ac.uk/people/paul/publications/RaysonEtAl_CL2007.pdf

Rayson, Paul, Dawn Archer, Scott Piao and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition: Semantic labelling for NLP tasks, in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 7–12. Lisbon, Portugal. http://www.lancaster.ac.uk/staff/rayson/publications/usas_lrec04ws.pdf

Rives, Amélie. 1888. *A brother to dragons and other old-time tales*. New York: Harper & Brothers.

Sanchez-Stockhammer, Christina. 2018. *English compounds and their spelling*. Cambridge: Cambridge University Press.

Schafer, Liz. (n.d.). A Shrew and The Shrew. British Library. Available online at https://www.bl.uk/treasures/shakespeare/shrew.html. Last accessed 11 March 2021.

Sharpe, James A. 1987. *Early Modern England: A social history, 1550–1750*. London: Edward Arnold.

Smith, Thomas. 1583. *DE REPVBLICA ANGLORVM. The maner of Gouernement or policie of the Realme of England*. London: Henry Middleton. Available online at http://name.umdl.umich.edu/A12533.0001.001. Last accessed 1 April 2021.

Taylor, Gary, John Jowett, Terri Bourus and Gabriel Egan. 2016. *The new Oxford Shakespeare: William Shakespeare, The complete works, Modern critical edition*. Oxford: Oxford University Press.

Wilson, Andrew and Jenny Thomas. 1997. Semantic annotation. In R. Garside, G. Leech and T. McEnery (eds.). *Corpus annotation*, 53–65. London: Routledge.

Wrightson, Keith. 1982. *English society, 1580–1680*. London: Hutchinson.

Wrightson, Keith. 1991. Estates, degrees, and sorts: Changing perceptions of society in Tudor and Stuart England. In P.J. Corfield (ed.). *Language, history and class*, 30–52. Oxford: Blackwell.