DRAFT

General Introduction to the Encyclopedia of Shakespeare's Language

1. The Encyclopedia

We celebrate Shakespeare because of what he created in language. It is strange then that a trip to any library will find shelves groaning with the vast quantity of literary criticism, but just a few volumes on his language. Moreover, in spite of 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. The AHRC-funded Encyclopedia of Shakespeare's Language Project (2016-2019) set out to fill this gap and to bring scholarship on Shakespeare's language fully into the 21st-century. The major output from project is the Encyclopedia. In the current phase of the project, the Encyclopedia consists of two volumes: one being a type of dictionary focusing on individual words, and the other being a compendium of semantic patterns focusing on groups of words.

Being corpus-based implies both a particular method for revealing meanings, and a particular theoretical approach to meaning. There is less reliance on the vagaries and biases of editors, and a greater focus on the evidence of *actual usage*. The question "what does X mean?" is pursued through another question: "how is X used?" We used computers to identify patterns of use across Shakespeare's works, some of which would be difficult for the human reader to encompass on such a large scale. An analogy might be aerial photography. Standing in a very large field, one might not be sure about the possible paths to take across it. But a photograph taken from an aircraft would reveal the paths that walkers have followed over the years. A computer can reveal frequency information that enables us to map regularities of use, and moreover, regularities in the contexts of use. We can see, for example, that the most regular use of the word *good* was as part of a polite address, such as *My good Lord*, said when meeting or parting from somebody, thanking them or making suggestions; or that the uses of the word *Irish* had overwhelmingly negative associations, except when it was applied to rugs (Irish wool was viewed positively). Moreover, we can track words across broader contexts of use, including the social characteristics of their speakers (e.g. whether they are male or female, high status or low status), and the kind of genre that characterises the play in which they are spoken (i.e. whether tragedy, comedy or history).

A striking feature of the Encyclopedia is that it also encompasses Shakespeare's language in the context of early modern English. It affords a sense of what Shakespeare's contemporary Elizabethan audience might have understood by his language. Current Shakespeare "dictionaries" focus squarely on language in the texts that Shakespeare is purported to have written. Even a check on the meanings in the Oxford English Dictionary does not take us very far outside Shakespeare, as the entries for that period are often heavily influenced by Shakespeare's language, leading to a degree of circularity. In contrast, the Encyclopedia compares Shakespeare's usage with 380 million words in texts across all extant genres, 1580 to 1640. This enables the discovery not only of specific usages characteristic of Shakespeare, but also the contemporary stylistic, discoursal and effects/attitudinal flavour of linguistic items (e.g. whether certain words were considered colloquial, religious, courteous, offensive, and so on).

The Encyclopedia is designed so that it appeals to both linguistic and literary Shakespearean interests and beyond. It will be of value to linguists with an interest in early modern English, as its remit encompasses the language of a major part of that period. It will be of value to literary scholars with an interest in early modern writers, as its remit encompasses the language not only of Shakespeare, but also that of his contemporaries. It will be relevant to many humanities disciplines – history, for example – as it will reveal contemporary attitudes, social constructions, and theatre scholars and professionals will find help in interpreting words, plays, characters and themes. The Encyclopedia's audience is not restricted to seasoned scholars, as part of its mission is to improve understandings of Shakespeare's language more generally. We have designed it so that it is relevant and accessible to university postgraduates and undergraduates, and senior school students.

The volumes of the Encyclopedia are available in paper and electronically (via Bloomsbury and associates' Drama Online website: http://www.dramaonlinelibrary.com/). A "lite", stripped down version of volume one will be available as an app for mobile phones, tablets and so on. In addition, the project has made publically available both its comprehensively annotated Shakespearean texts and its work on the comparative data. Access is largely through the web-based corpus analysis interface CQPweb, which is powerful yet relatively user-friendly. The point of doing this is that it will allow people to replicate what we did or, more particularly, use similar techniques to do what we might not have thought of doing. Details about the datasets, tutorials on how to use Lancaster University's CQPweb system, further publications on Shakespeare's language (e.g. papers, presentations and blogs), and more can be found on the project's website (http://wp.lancs.ac.uk/shakespearelang/).

2. The sources of Shakespeare's language

The core data

What does the label "Shakespeare's language" actually refer to? Clearly, we are not talking about the bulk of the language he produced in his lifetime, namely, his everyday spoken language, of which we have no record. "Language", then, refers to the language of his literary output. What about the first half of the label, "Shakespeare's"? Today's authors generally write alone. In the early modern world, collaboration amongst playwrights is known to have been very common. Indeed, plagiarism is a modern notion; in the early modern world re-using portions of text from the works of others was, in some cases, construed as complimentary. Furthermore, unlike today's authors, Shakespeare would have had no clear authorial oversight of his works. 36 plays were put together, 20 of which had not been published before, and published as the First Folio in 1623, that is, 7 years after he died. There was thus no authorial input into copyediting or proof correcting (even if those processes had existed). Around 18 plays had been previously published as quartos, but some have been considered "bad" (perhaps reconstructed from memory). In any case, and more generally, the whole notion of publishing plays in early modern England was rather different from what it is today. An early play-text was a bundle of manuscript fragments written for performance, rather than a unitary whole written for publication. In the light of all this, it is best to consider the label "Shakespeare's language" as a shorthand for surviving written literary texts that purport to represent, for the most part, language that Shakespeare produced.

To achieve the goals of the Encyclopedia, the project needed one stable body of data to examine. This is particularly a consequence of our method. For example, if you want to compare across frequencies of items in Shakespeare with those of his contemporaries, it is essential to base the Shakespeare frequencies on the same body of data. The largest single body of Shakespeare's works and the earliest publication of a large group of his works is that constituted by the First Folio (1623). This was the obvious choice for the Encyclopedia. Needless to say, scholars have recognised the presence of other hands in plays listed in the First Folio, despite the attribution in its title: *Mr William Shakespeare's Comedies, Histories and Tragedies*. For example, it is thought that Marlowe, Nashe, Peele and Greene contributed to the *Henry VI* plays; Middleton's hand is seen in *Measure for Measure, Timon of Athens* and *Macbeth* and John Fletcher co-authored *Henry VIII or All Is True*. Collaboration works both ways, of course, and there are many claims of Shakespeare's hand in plays *not* in the First Folio; famously, one manuscript page of *The History of Thomas More* is believed to be an example of Shakespeare's handwriting. To have included 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 Encyclopedia project in debates about authorship attribution before it had even got off the ground. Consequently, we limited ourselves to texts with a relatively longstanding scholarly acceptance into the Shakespeare canon, the First Folio texts, and added two further plays to our core data: *Pericles* (Quarto 1) and *The Two Noble Kinsmen* (Quarto 1), believed to be collaborations with George Wilkins and John Fletcher, respectively. To be clear, this decision does not constitute a claim on our part that the First Folio is in some sense more "Shakespearean" than other texts, such as the Quartos, or that is was somehow immune to the normal text production practices of the day (e.g. the less than consistent practices of compositors).

Table 1 lists the plays that constitute our core Shakespeare Corpus, as described above. Importantly, it includes additional information about those plays. It displays: a short title for each play included in the core data; an abbreviation for each (following those used by *The Arden Shakespeare*); the play genre, that is, whether it was considered a tragedy, comedy or history (following designations given in the First Folio itself, with the exception of *Cymbeline* which is re-classified from a tragedy to a comedy, according to convention in modern editions); and probable dates of first publication and production.

Table 1. The plays constituting the core Shakespeare Corpus

Play (short title)	*Abbreviation*	*Genre (tragedy, comedy, history)*	*Date of first publication*	*Date range of 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	1594	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	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	TNK	C	1634	1613-1614	1613

Some of the information in Table 1 is not without its controversies. The notions of tragedy, comedy and even history are slippery, and some of the Folio designations are puzzling. Cases in point include the so-called "problem plays", which are conventionally taken to be *All's Well That Ends Well*, *Measure for Measure* and *Troilus and Cressida*. These combine elements of comedy and tragedy. We tracked these three plays as a group, in order to allow their specific characteristics to be revealed. There is fairly good evidence for the first publication dates for Shakespeare's plays, but often thin evidence for the date of first production. This is why the table supplies a date range in which the first production probably occurred, and an approximate specific date – a date which should be taken as a best guess. All first production dates are drawn from *DEEP: Database of Early English Playbooks* (edited by Alan B. Farmer and Zachary Lesser; http://deep.sas.upenn.edu/).

Subsidiary data

Whilst our core basis for the analysis of Shakespeare's language is as described above, we did accommodate further works as subsidiary datasets. Some of Shakespeare's plays were published in quarto before they appeared in folio. 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. In order to check whether whatever we were saying on the basis of the First Folio would differ from what might emerge from the Quartos, we created a collection of 22 Quarto plays consisting of the following:

Hamlet, First Quarto (1603); *Hamlet, Second Quarto* (1604); *Henry IV Part 1, Quarto 0* (1598); *Henry IV Part 1, Quarto* 1 (1598); *Henry IV Part 2, Quarto 1* (1598); *Henry V, Quarto 1* (1600); *Henry VI Part 2, Quarto 1* (1594); *Henry VI Part 3, Octavo 1* (1595); *King Lear, Quarto 1* (1608); *King Lear, Quarto 2* (1619); *Love's Labour's Lost, Quarto 1* (1598); *The Merchant of Venice, Quarto 1* (1600); *The Merry Wives of Windsor, Quarto 1* (1602); *A Midsummer Night's Dream, Quarto 1* (1600); *Much Ado About Nothing, Quarto 1* (1600); *Othello, Quarto 1* (1622); *Richard II, Quarto 1* (1597); *Richard III, Quarto 1* (1597); *Romeo and Juliet, Quarto 1* (1597); *Romeo and Juliet, Quarto 2* (1599); *Titus Andronicus, Quarto 1* (1594); *Troilus and Cressida, Quarto 1* (1609)

Thus far, we have only mentioned Shakespeare's plays. Shakespeare's poems are especially dense in technical problems for corpus-based analysis. Aside from the minor matter that the texts are structured differently, they are particularly challenging for some of our semi-automated analyses, notably grammatical part-of-speech tagging. This is because those analyses partly work on the basis of probabilities about (a) what part-of-speech a word normally is, and (b) what part-of-speech a word normally is given the parts-of-speech of the words surrounding it. But those norms are precisely the norms that poetry often violates for effect. Consequently, we treated the poems in the same way as the Quartos, namely, we created a subsidiary dataset containing the following:

A Lover's Complaint Quarto (1609); *The Passionate Pilgrim Octavo* (1599); *The Phoenix and the Turtle Quarto* (1601); *The Rape of Lucrece Quarto* (1594); *The Sonnets Quarto 1* (1609); *Venus and Adonis Quarto 1* (1593)

3. The base texts: Form, transcription and structure

The path of least resistance for the project would have been to have used an edited, modernised edition. There are, however, three reasons why we did not take this path. One is that modern editions of Shakespeare's works are often collations of the First Folio and Quarto texts, edited to appeal to a variety of audiences (but not usually linguists). They are a kind of reimagined historical text, but not actually a historical text that can act as an anchorage point. Another is that modern editors, with their different priorities, often strip out the very things that might interest a linguist, such as early modern verb or noun endings. And the final reason is that 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 versions of Shakespeare's texts (we have ignored the further complication of spelling variation here). For the reader, of course, how compounds are represented is hardly an impediment. But for computer analysis, some acute problems are raised. A policy of closing such compounds will produce a very different word count from one of opening them (remember that our analyses are partly driven by statistical regularities). For these reasons, we opted for original spelling transcriptions. That way, we could control consistency. Note that when we refer to consistency here, we refer to consistency in the way the original text is transcribed and made ready for computational linguistic analysis (we are not making claims about the internal consistency of the First Folio).

For language to be computer searchable, text needs to be in digitized form. The Encyclopedia project's electronic files for all of the works listed in this section were kindly provided by Internet Shakespeare Editions (ISE) (http://internetshakespeare.uvic.ca/). The ISE, as stated on their website, "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" (http://isebeta.uvic.ca/Foyer/quality/). Crucially, the ISE were able to provide us with original spelling versions of Shakespeare's works. These transcriptions were initially based on Charlton Hinman's *The Norton Facsimile: The First Folio of Shakespeare* (1968), but then also electronically checked against transcriptions held in the Oxford Text Archive, with appropriate corrections made at that point. ISE state that their "old-spelling versions are diplomatic transcriptions and do not amend any errors present in the original text" (http://internetshakespeare.uvic.ca/Foyer/plays). Diplomatic transcriptions are faithful warts-and-all transcriptions, which suited our aim of building on an original, untampered foundation. However, the idea of a faithful transcription of an extensive historical document is more of an ideal than a reality. Consequently, although we did not systematically proofread the transcription against the original, we did investigate, and if necessary correct, any oddities in the transcription. 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.

The ISE files were in XML format. This has become the norm for digital editions, corpora and so on. Essentially, XML (Extensible Markup Language) involves a standardised way of describing texts or data, both in terms of their structure and their contents. The building blocks or elements of texts are marked off by pairs of angle brackets in which codes indicate the nature of those blocks. The building blocks of Shakespeare's plays typically include the plays, acts, scenes and speeches of particular characters. As that sequence suggests, blocks can be nested inside blocks. Fortunately, the vast bulk of the XML formatting required was already in place in the ISE files, though we checked it and made some adjustments to suit our own needs.

4. Regularisation of spelling

A general consideration for any computer-aided analysis is that the linguistic item you may be searching for could have more than one form, even in present-day data. The most obvious example of this is morphological variation. Thus, searching for *love* will not retrieve all instances of *loves*, *loving* and *loved*. Yet all of these belong to the same word family, or what linguists refer to as lexeme. In Shakespeare's texts, one finds further variation in the shape of archaic forms, such as the noun plural marker *-(e)n* (e.g. *shoen* "shoes"; *eyen* "eyes"). In addition, a key consideration when it comes to historical texts is spelling variation, as spelling standardisation was not largely complete until towards the end of the 17th century. Using a program to find all examples of *sweet* would miss *sweete*, *love* would miss *loue*, and *doubt* would miss *doute*, and some spellings are ambiguous (e.g. in early texts, *than* could either be today's *than* or *then*). Our solution was to use the program *Variant Detector* (VARD), developed by scholars at Lancaster University over more than 15 years, and most significantly by Alistair Baron (see http://ucrel.lancs.ac.uk/vard/about). This program regularises variation by matching variants to 'normalised' equivalents using a search and replace script, as well as contextual information to tackle ambiguities and an additional lexicon to treat word forms that are specific to or have undergone semantic change since the early modern period. It is crucial to note that the program does not delete the original spelling, but places it in a specific XML element, thereby making it easily available for inspection (frequent spelling variants found in the First Folio are given in Volume 1). Regularisation here means to link one spelling with another, one usually being less regular than the other. Because the project demanded a high level of accuracy, we did not run the program in fully automatic (whole-text) mode. Instead, the program's manual (word-by-word) mode can on most occasions suggest regularisation options in order of likelihood, from which the human operator approves a selection. We made no attempt to "correct" the spelling, with very rare exceptions made for obvious printer errors, such as *aud* for *and*.

What does one regularise the spelling to? There was no standardised spelling in the way that there is today. Our general policy was to:

- Preserve morphology, e.g. second and third person verb inflections (*-(e)st*, *-(e)th*), past tense forms (e.g. *holp*), past participle forms (e.g. *holpen*), plural forms (e.g. *shooen*), non-standard superlatives (e.g. *horrider*), and *you*/*thou*;
- Preserve obsolete, archaic or rare forms, e.g. *cozen/ed*, *haply*, *morrow*;

- Only use a form that had currency in Shakespeare's time;
- Prioritize the most frequent spelling in Shakespeare's work; and
- Leave any ambiguous or indeterminate cases as they are.

We also made some changes so that groups of word-forms could be better matched (e.g. as lexemes) or grammatical features better detected. These included changes to apostrophes (e.g. *do's* → *does*; *the Dukes table* → *the Duke's table*) and some expansions (e.g. *qd* → *quod*). Regarding compounds, there is no perfect solution. An important consideration here is where in the Encyclopedia the compound might appear, if indeed it is to appear at all. For example, regarding the open compound *to morrow*, if left unchanged, each of the two components would be treated separately under the entries for *to* and *morrow*, rather than together under the entry for *tomorrow*. Consequently, we joined (and in the case of closed compounds split) such items according to the conventional compound form in present-day English. For hyphenated compounds, where there is much less clarity even in present-day English, we followed *The Arden Shakespeare*, unless their form was unusual or required the introduction of hyphens where there is a closed form in present-day English. Thus, for example, *first born* became *first-born*, but *howerglasse* became *hourglass* (not *hour-glass*, as in Arden). It is worth reiterating that any changes we made did not involve the loss of the original: all is preserved by the program VARD in the XML. Analysis of all of the regularisation patterns in VARD can be achieved by another program written by Alistair Baron, DICER (Discovery and Investigation of Character Edit Rules).

The above by no means exhausts all the issues that attended the regularisation of spelling in Shakespeare's works. What about dialectal words? These notably include words whose spellings seem to indicate a particular accent, such as the regular substitution of the letter *b* for the letter *p* in speeches for Welsh characters (e.g. Fluellen's *pig* for *big*). We linked these to the regular forms (as identified along the lines indicated in the previous paragraph). This ensured, for example, that *pig* for *big* did not appear as a separate headword in volume 1 but could be treated under the entry for *big*. We also coded dialectal items according to the following regions: English, Irish, Welsh, Scottish and Foreign. Foreign dialect was used to mark the spellings of English words indicating non-English pronunciation, such as French-accented English words (e.g. *dat* for *that*; many examples can be found in *Henry V*, 5.2). What about non-English words? Words that are distinctly foreign were coded as French, Latin, Spanish, Italian or (Other) Foreign language. Such words were also regularised according to the spelling given in *The Arden Shakespeare*. Thus, *adiew* was regularised to *adieu* and marked as French. On occasion, the 'foreign' status of word was not entirely clear – was it really foreign or had it become naturalised as an English loanword? In such cases, we made a judgement based on its contexts of use, not just in Shakespeare but also more broadly in early modern English, and especially depending on whether it generally appeared in the context of French texts or speakers. *Monsieur* is a case in point. In early modern English, it was used most frequently in interactions with French people, and so we coded it French.

A final area to comment on concerns so-called corruptions and misuses, including malapropisms. These were dealt with on a case-by-case basis, using the kinds of principles outlined above. *Coram*, for example, was a common corruption of Latin *quorum*. It was thus marked as Latin and linked to the regular form *quorum*. *Fartuous* is one of Mistress Quickly's famous malapropisms (*The Merry Wives of Windsor*, 2.2), and was linked to its regular form *virtuous*.

5. Grammar: Parts of speech

Texts or a corpus can be enhanced by the addition of interpretative information in the form of annotations or codes. Grammatical annotation means that grammatical information has been added; linguistic items have been associated with particular grammatical categories. Such annotation can be useful in a situation where it is useful to be able to distinguish the grammatical functions of particular word form. For example, *love* might be represented as *love_NN1*, in which case the tag appended to the word by an underscore indicates that the word is a singular common noun. Had the tag been _VV0, it would have indicated that the word is a verb in its base form. A set of commonly used grammatical tags, the CLAWS6 tagset, can be found at: http://ucrel.lancs.ac.uk/claws6tags.html. This tagset formed the basis of our grammatical analyses. Such grammatical matters may seem the sole preserve of the grammarian, but for the lexicographer they are also in fact crucial, as *love* the noun can receive a separate entry from *love* the verb. Word-meanings develop in conjunction with grammar, and it makes sense to list words – strictly speaking lemmas or headwords – separately according to their parts-of-speech. Furthermore, more generally, grammatical information also enables richer analyses of patterns of meaning.

Part-of-speech tagging software has been under development at Lancaster University since the early 1980s. The software is called CLAWS (the Constituent Likelihood Automatic Word-tagging System: see http://ucrel.lancs.ac.uk/claws). In a nutshell, CLAWS works on the basis of (1) a lexicon, including words (or multi-word units) and suffixes and their possible parts of speech, and (2) a matrix containing sequencing probabilities (e.g. the likelihood that the word following an adjective will be a noun), which is applied to each sentence to disambiguate words which could potentially be several parts-of-speech. CLAWS is claimed to achieve 96-97% accuracy on written texts, and a slightly lesser degree of accuracy on spoken texts. However, the spelling variability in early modern English texts presents a problem for automated tagging. Due to the presence of variants, fewer words can be matched in the CLAWS lexicon, and the tagger thus has to make a best guess about words which it ought to have been able to recognise. By regularising the spelling of our texts using VARD, we solve part of this problem, but not all of it. Even if we had perfect regularisation, there is still the issue of vocabulary change over time. Some words have disappeared from English over the last 400 years (e.g. the verb *wot* "know" or the adverb *iwis* "certainly") and are thus not in the tagger's lexicon. Others still exist but behave differently in grammatical terms (e.g. *fee* is only a noun in contemporary English but could equally well be a verb in early modern English). A study that three of the project team members conducted some years ago showed that, even with regularised Shakespeare texts, CLAWS only achieved 89% accuracy. That might sound quite good, but remember that this would be roughly equivalent to 1 error in every 10 words – certainly not good enough for the Encyclopedia. A second problem with using CLAWS was that its design for contemporary English means that it overlooks many grammatical features of early modern English – for instance, the existence of *thou* and *thee* as forms distinct from *you*, rather than as marginal phenomena as they are today; or the regular use of an inflected second person for all verbs, including the otherwise non-inflecting modal verbs (e.g. *mayst*, *canst*). Our solution to both these problems was to make adjustments to CLAWS, and also to manually check all core Shakespeare texts.

The strategy we adopted was to make changes and additions to the CLAWS lexicon and tagset, while using the same probability matrix that was developed for contemporary written English. The kinds of grammatical cues that the matrix encodes (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 grammar which have changed *least* over the past 400 years. By contrast, the things that *have* changed – as indicated above – are individual words and phrases and their particular grammatical behaviour. We were able to modify the CLAWS lexicon so that it covered around 3,300 additional word-forms (most of these sourced directly from the regularised First Folio) and around 80 additional phrases, as well as changing the probability profiles of another approximately 80 common words. We also added extra tags for the second-person singular, which were overlaid onto the output after the completion of normal CLAWS processing. Finally, we enhanced CLAWS' ability to separate out pronoun-verb contractions. While many of these still exist today and are accurately handled by CLAWS already (e.g. *he's*, *she'll*, *I'm*, *you've*), there were more of these in Shakespeare's day (e.g. *methinks*, *twill*) and for the sake of consistency they also had to be broken apart into separate word units.

For the First Folio dataset, we applied a further layer of analysis to assure the accuracy of the grammatical classifications: *manual post-editing*. This is exactly what it sounds like: the entirety of each tagged text is inspected by a human trained in the tagging system, checking each word one at a time for correctness, and correcting all the mistakes as they were identified. This kind of post-editing does not completely guarantee accuracy of tagging, since human error is always a factor – and there sometimes occur constructions on which different human analysts might disagree. Rather, we may state that the grammatical tagging in our First Folio data is as close to 100% accuracy as it is humanly possible to be. A fast post-editor can check the tagging of a Shakespeare play in about the time it would take to read that same play carefully. This investment of time was clearly worthwhile for the Encyclopedia's core dataset, but it was not feasible to apply it to the other, much larger ancillary datasets discussed below; these, as well as the Shakespeare Quartos were only processed by the adapted CLAWS software, without post-editing.

6. The categorisation of social features

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. In the fictional worlds of Shakespeare's plays, we can investigate the social construction of characters. This is pertinent to Volume 2, which has a particular focus on characters (for more detail on how we enacted social categories, see the Introduction to Volume 2). But there is also relevance for Volume 1. The regular choice of words to construct particular identities (male, female, high-ranking, etc.) may give us clues about the meanings of those words – they may have associations of femaleness, high status, and so on.

Categorising characters as male or female is relatively straightforward, though it was necessary to develop separate categories for characters with an assumed identity (e.g. a female character playing a male character) (see the Introduction to Volume 2). The categorisation scheme to capture a character's status/social rank draws upon the approach developed by Archer and Culpeper (2003). Categories are designed to reflect the pre-industrialised nature of early modern society, something which required us to review the relevant work of historians and historical linguists.[1] They are also designed to reflect the way in which early modern contemporaries spoke about status. Commentators like Sir Thomas Smith (1583) pointed out that men [sic] could be divided into *foure sortes* during Shakespeare's time, namely, gentlemen, citizens, yeoman artificers and labourers. Thanks to the work of historians, we know that the population of England grew from just over 3 million to just over 4 million during Elizabeth I's reign (1558-1603). We also know that estimates for the number of gentry vary from 15,000 to 20,000. This allows us to assume that the vast majority of the population belonged to classes below the gentry. As such, we have been careful to adopt a categorisation

[1] See, for example, Holmes (1982), Wrightson (1982, 1991), Sharpe (1987), Corfield (1995), Hunt (1996), Archer and Culpeper (2003), and Innes (2007).

scheme which can distinguish gentry from professionals and other middling groups, as well as distinguish ordinary commoners from the lowest groups.

Our categories are shown in Table 2, along with the numerical code used to assign them, a brief explanation and prototypical examples.

Table 2. Social status categories

Social category	*Numerical code*	*Description*
Monarchy	0	The rulers of subjects. Prototypical examples – King, Queen, Majesty.
Nobility	1	Those with particular inherited or conferred ‘titles’ that allow them to sit in the House of Lords, including the Lords ‘spiritual’. 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, which is 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, to use Addison’s (1711) phrase, the ‘three great professions’ of Law, Medicine and the Church. Prototypical examples – clergymen, lawyers, medical practitioners, school-teachers, military.
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 – manufacturers, wholesalers, retailers, merchants, money-lenders, skilled craftsmen, and financiers. Prototypical examples – merchant, shopkeeper, carpenter, shipbuilder, warehouseman, cloth dealer.
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 – ‘labouring folk’, yeomen, poor husbandmen, wage labourers, apprentices to the non-professional occupations.
Lowest Groups	6	Common seamen, servants, cottagers and paupers, the unemployed, common soldiers and vagrants. Prototypical examples – servant, vagrant.
Supernatural Beings	7	Prototypical examples – ghost, fairy, god, sprite, apparition.
Problematic	P	Those whose status was uncertain at the time: e.g., actors, and characters who undergo a significant change in status during the play.

Six categories – Gentry, Professional, Other Middling Groups, Ordinary Commoners and Lowest Groups – have been adopted without change from Archer and Culpeper’s (2003) system. The Nobility category has been adapted, to enable us to separate Monarchy from Nobility in our system. 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. We are sensitive to the fact that we are working with fictional data, hence, for example, our addition of a Supernatural Beings category accounting for the forty-plus ghosts, god, fairies, etc. in the thirty-eight plays.

7. Comparative Corpus of Playwrights

The comparative data, briefly mentioned in section 1, constitutes a valuable dimension to our Encyclopedia by enabling us to set Shakespeare’s language into the wider dramatic and linguistic landscape of his day.

The first of two comparative sets of data we used was a specialised corpus of other early modern English plays designed to be as close as possible in size, date and proportions of comedy, history and tragedy genres to our core Shakespeare dataset of 38 plays (detailed in section 2). It comprised 46 plays by 24 playwrights, with first production dates spanning the period 1584-1626, and was compiled with expert guidance from the project's Renaissance drama specialist. A balance of relatively early and late plays were included to reflect those by Shakespeare, which are generally considered to have undergone language style changes over the period in which he was writing. In addition to proportioning the amounts of comedy, history and tragedy plays in line with the Shakespeare dataset, we also paid attention to sub-genres, ensuring, for example, that domestic and revenge tragedy, and romantic, pastoral and tragi-comedy were all represented. We included some city comedy in the corpus of comparative play-texts because it was a popular style of the period, although it was not one which Shakespeare himself favoured. We balanced the proportions of dialogue spoken by male and female characters to approximately that of the Shakespeare dataset, with some difficulty (female dialogue tends to be scarcer in plays by his contemporaries). However, we included plays by male authors only, because the Shakespeare data is male-authored and comparisons of authors of different sex were beyond the project's remit. We did not exclude collaborative plays if they otherwise fitted our dating and genre criteria, because (as noted in section 2), it was common in playwriting of the period and some of the plays in the Shakespeare canon are known to be collaborative to some extent. However, we minimised the inclusion of plays by anonymous authors, especially any considered to be part of the 'apocrypha' of plays in which Shakespeare may have had a hand in writing.

The play-texts were sourced in a digital format from the Early English Books Online Text Creation Partnership (EEBO-TCP) (http://www.textcreationpartnership.org/tcp-eebo/), and are listed in the table below.

Table 3. The plays constituting the Comparative Corpus of Playwrights

Play (short title)	***Abbreviation***	***Author(s)***	***Genre (tragedy, comedy, history)***	***Date of first publication***	***Date range of first production***	***Approx. date of first production***
The Spanish Tragedy	CCTSPANT	Kyd, Thomas	T	1592	1585-1589	1587
The Jew of Malta	CCTJEWOF	Marlowe, Christopher	T	1633	1589-1590	1589
Dr Faustus	CCTFAUST	Marlowe, Christopher	T	1604	1592-1593	1592
Dido Queen of Carthage	CCTDIDOC	Marlowe, Christopher	T	1594	1585-1586 or 1591	1586
The Malcontent	CCTMALCO	Marston, John	T	1604	1602-1604	1604
A Woman Killed With Kindness	CCTAWKWK	Heywood, Thomas	T	1607	1603	1603
Sejanus	CCTSEJAN	Jonson, Ben	T	1616	1604?	1604
The Maid's Tragedy	CCTMAIDT	Beaumont, Francis	T	1619	1610-1611	1610
The White Devil	CCTWHITE	Webster, John	T	1612	1612-1613	1612
The Duchess of Malfi	CCTDOFMA	Webster, John	T	1623	1612-1614	1614
The Changeling	CCTCHANG	Middleton, Thomas	T	1653	1622	1622
Women Beware Women	CCTWBEWA	Middleton, Thomas	T	1657	1613-1621	1621

The Scottish History of James the Fourth	CCHJAMES	Greene, Robert	H	1598	1590	1598
Tamburlaine Part I	CCHTAMB1	Marlowe, Christopher	H	1590	1587-1588	1587
Edward II	CCHEDWII	Marlowe, Christopher	H	1594	1591-1593	1592
Edward I	CCHEDWAI	Peele, George	H	1593	1590-1593	1591
The Massacre at Paris	CCHPARIS	Marlowe, Christopher	H	1594	1593	1593
The Battle of Alcazar	CCHALCAZ	Peele, George	H	1594	1588-1589	1589
The Death of Robert, Earl of Huntingdon	CCHDEATH	Munday, Anthony	H	1601	1598	1598
Edward IV Part I	CCHEDIV1	Heywood, Thomas	H	1600	1592-1599	1599
Edward IV Part 2	CCHEDIV2	Heywood, Thomas	H	1600	1592-1599	1599
Sir John Oldcastle	CCHOLDCA	Munday, Anthony	H	1600	1599	1599
If You Know Not Me, You Know Nobody Part I	CCHIFYO1	Heywood, Thomas	H	1605	1604-1605	1604
Sir Thomas Wyatt	CCHWYATT	Dekker, Thomas	H	1607	1602	1602
The Valiant Welshman	CCHWELSH	R. A., Gent.	H	1615	1610-1615	1612
The Duchess of Suffolk	CCHDUCHE	Drue, Thomas	H	1631	1624	1624
Alexander and Campaspe	CCCALEXA	Lyly, John	C	1584		1583
Gallathea	CCCGALLA	Lyly, John	C	1592	1583-1585	1585
Friar Bacon and Friar Bungay	CCCFRIAR	Greene, Robert	C	1594	1589	1586-1590
The Old Wives' Tale	CCCOLDWI	Peele, George	C	1595	1588-1594	1590
The Blind Beggar of Alexandria	CCCBLIND	Chapman, George	C	1598	1596	1596
The Fair Maid of the West Part I	CCCFAIR1	Heywood, Thomas	C	1631	1597-1604	1604
An Humorous Day's Mirth	CCCANHUM	Chapman, George	C	1599	1597	1597
Two Angry Women of Abington	CCCTWOAN	Porter, Henry	C	1599	1598?	1598
Mucedorus	CCCMUCED	Anon.	C	1598	1588-1598	1590
Old Fortunatus	CCCOLDFO	Dekker, Thomas	C	1600	1599	1599
How A Man May Choose	CCCCHUSE	Heywood, Thomas	C	1602	1601-1602	1602
Volpone	CCCVOLPO	Jonson, Ben	C	1616	1605-1606	1606
The Woman Hater	CCCHATER	Beaumont, Francis	C	1607	1606	1606
The Miseries of Enforced Marriage	CCCMISER	Wilkins, George	C	1607	1605-1606	1606
The Faithful Shepherdess	CCCFAITH	Fletcher, John	C	1610	1608-1609	1608
The Roaring Girl	CCCROARI	Middleton, Thomas	C	1611	1611	1611

The Knight of the Burning Pestle	CCCKOFBP	Beaumont, Francis	C	1613	1607	1607
Philaster	CCTPHILA	Beaumont, Francis	C	1620	1609	1609
Bartholomew Fair	CCCBFAIR	Jonson, Ben	C	1631	1614	1614
The Bondman	CCCBONDM	Massinger, Philip	C	1624	1623	1623

As with the Shakespeare data, all first production dates are drawn from *DEEP: Database of Early English Playbooks* (edited by Alan B. Farmer and Zachary Lesser; http://deep.sas.upenn.edu/).

Once downloaded, the digitised play-texts were carefully checked, and any missing or unclear text corrected as far as possible by cross-referencing with the facsimile printed manuscript files (also on EEBO-TCP). The play-texts were XML-tagged in a similar way to the Shakespeare dataset (described in section 3), and the spelling was normalised using the VARD software, deployed in its automatic mode with benefit of being 'trained' manually on the Shakespeare data (detailed in section 4). The characters in the comparative corpus of plays were also assigned categories for social status, using the framework explained in section 6.

8. Comparative Corpus of Early English Books Online (EEBO) and genres

The second of the two comparative sets of data we used is much larger. Thanks to the rise of Digital Humanities, we now have electronic 'big data' available to source a comparison point for Shakespeare's works. By observing such data through the corpus linguistic lens, we can achieve 'macro readings' of thousands of texts, which makes comparisons possible on a scale that would not be achievable otherwise. The world of scholarship has recently seen the arrival of a transcribed version of Early English Books Online – Text Creation Partnership (EEBO-TCP), of which 380 million words span the 80-year period 1560-1639. This allows us to not only compare Shakespeare's linguistic usage with that of his contemporaries, but also to examine what his contemporaries thought the language meant. We can tap into those broad patterns of meaning and gain a sense of what was triggered in the minds of the Elizabethan audience when they heard Shakespeare's words. We will also cross-check some of those word-meanings with what contemporaries wrote about them in early dictionaries and glossaries, using the historical data base *Lexicons of Early Modern English* (LEME) (http://leme.library.utoronto.ca/).

As we already mentioned in section 1, our aim is not simply to reveal the denotative or conceptual meanings of words and other linguistic units, but also something of their social and stylistic 'flavour' in the general language of the period. The comparative corpus of playwrights, outlined in section 6, has the annotation required for social comparisons. What we are particularly looking for in the larger EEBO-based corpus is stylistic flavour. In order to achieve this, we needed to examine, for example, the dispersion of words, that is, whether they clustered in particular genres or registers where they might acquire their particular stylistic flavour (thus, for example, legal-sounding language would cluster in legal genres). To achieve this, we had to enhance the EEBO-TCP data by designing and applying a genre classification scheme to some 5,900 texts from the 1560-1639 period, texts which encompass literature, philosophy, politics, religion, geography, science, law and many other fields. Preliminary work on assigning genres and sub-genres to these texts, largely based on self-description in titles (e.g. *play, sermon*, *chronicle*, *treatise*, etc.) had already been carried out by Lancaster scholar Tony McEnery. Our task was to design a classification system to group McEnery's genres and sub-genres according to domain and style, while providing conceptual and stylistic labels that are accessible to readers of the Encyclopedia. The resulting classification is shown in Table 4. We applied this scheme to each of the 5,900 texts, removing any texts duplicated in our core Shakespeare Corpus (Table 1) or Comparative Corpus of Playwrights (Table 3). We acknowledge that a degree of fuzziness and overlap amongst categories remains. Whilst some were relatively easy to separate out and place, others had rather mixed contents and / or membership claims to multiple superordinate categories. In such cases, we made a judgement about best fit.

Table 4. Genre classification of EEBO texts for the period 1560-1639

Style	Domain	Genre	Sub-genres
Literary	Imaginative	Plays	Comedy, History, Tragedy, Masque
		Poetry, Verse & Song	Ballads, Songs
		Fiction	
		General	
Formal – Spiritual	Religion	Bible	
		Catholicism	Anti-Catholicism
		Protestantism	Church of England, Church of Scotland, Non-Conformism
		Doctrine, Theology and Governance	Heresy, Prayer, Sin and Repentance,
		General	Articles, Christians, Devotional, Epistles, Sermons, Others
Formal - Statutory	Government	Royal	Communications and Orders, Proceedings
		Parliamentary	General, Proceedings and Reports
		Legal	Legislation and Orders, Trials and Disputes
		General	Declarations, Military, Proceedings, Speeches
Formal - Instructional	Didactic	Astronomy	
		Philosophy	
		Science	Experiments
		Mathematics	
		Medicine	Anatomy
		General	Alchemy, Almanack, Astrology and Predictions, Lecture
Informational	Factual	Biography	
		Colonial	
		Essay	Admonition, Advisory, Apologia, Argumentative, Commentary On People And Places, Death, Obituaries and Epigraphs, Dialogue, Exhortation, General Lamentations
		Letters	
		Pamphlets	Analysis And Instruction, Chronology, Directory, Finance and Trade, Food and Cookery, History, Language, Travel, Treatise, London, Petitions, Reportage, Satire, Wit and Humour
		General	