sciendo

ICAME Journal, Volume 44, 2020, DOI: 10.2478/icame-2020-0002

Issues and challenges in compiling a corpus of Early Modern English plays for comparison with those of William Shakespeare

Jane Demmen, Lancaster University

Abstract

In this article I discuss the issues and challenges of compiling a corpus of historical plays by a range of playwrights that is highly suitable for use in comparative, corpus-based research into language style in Shakespeare's plays. In discussing sources for digitised historical play-texts and criteria for making a selection for the present study, I argue that not just any set of Early Modern English plays constitutes a suitable basis upon which to make reliable claims about language style in Shakespeare's plays relative to those of his peers. I point out factors outside of authorial choice which potentially have bearing on language style, such as sub-genre features and change over time. I also highlight some particular difficulties in compiling a corpus of historical texts, notably dating and spelling variation, and I explain how these were addressed. The corpus detailed in this article extends the prospects for investigating Shakespeare's language style by providing a context into which it can be set and, as I indicate, is a valuable new publicly accessible resource for future research.

1 Introduction

In this article I discuss the issues and challenges encountered in the construction of a corpus of Early Modern English[1] (EModE) play-texts[2] by a range of playwrights contemporaneous with Shakespeare (whose plays were written between approximately 1590 and 1613). This corpus was compiled to meet the goals of a project which helps address the current gap in comparative corpus-based research into language style in Shakespeare's work, and which provides new resources for the study of historical drama: the Encyclopedia of Shakespeare's Language Project (ESLP).[3]

Plays by William Shakespeare have a unique place in English language and literature. Interest in and appreciation of them has been sustained over several

hundred years, and the investigation of language style in Shakespeare's plays using corpus linguistic methods is now well established. Existing studies range from close comparisons of individual characters (e.g. Culpeper 2009; Archer and Bousfield 2010) to analyses of language taking in all of Shakespeare's plays (for example, Ulrich Busse's 2002 investigation of second-person pronouns; Beatrix Busse's 2006 study of vocatives). Such work usefully complements the vast, longstanding body of mainly qualitative literary critical research by providing some quantitatively-based perspectives. However, Shakespeare was just one among a number of well-known and successful playwrights writing in the late 16th and early 17th centuries. In the field of corpus linguistics that explores language style features (corpus stylistics), relatively little attention has thus far been given to investigating the language of playwrights other than Shakespeare. Notable exceptions are Hope and Witmore (2010: 387–390) and Culpeper (2011), who emphasise the need for further comparative corpus-based research. Quantitatively-based comparisons of Shakespeare's language style relative to those of his peers are to date mainly restricted to computational stylistic research, focusing on authorship attribution (for example, the studies in Craig and Kinney 2009). Areas as yet largely unexplored include pragmatic phenomena, style features of dramatic sub-genres, metaphor use, and characterisation of people of different gender and social rank through language style. Corpus stylistic comparative studies such as the ESLP potentially help address this research gap.

The ESLP aims to exploit electronic corpora and corpus linguistic methods to provide new, empirically-based insights into Shakespeare's plays. This is achieved by situating Shakespeare's language in the context of language used in plays by other contemporaneous playwrights, and examining it at multiple levels (including words, phrases, semantic themes, profiles of characters and plays). To facilitate its aims, the project employs quantitative data as the basis from which to reveal which language style features used by Shakespeare are also used more widely in plays of the period, and in EModE more generally. To meet the above aims, three corpora were compiled which together form the Enhanced Shakespearean Corpus (ESC), one of the main project outputs. The ESC comprises, firstly, a main or target corpus of play-texts wholly or substantially attributed to Shakespeare (the ESC: First Folio Plus), which was compiled first. Subsequently, two reference corpora (that is, two other datasets which can be compared with the ESC: First Folio Plus on a statistical basis using corpus linguistic software tools) were compiled to facilitate the extraction of comparative, quantitative data. One is a corpus of play-texts by other contemporaneous playwrights that is similar in size and structure to the ESC: First Folio Plus, the

ESC: Comparative Plays, which is the focus of this paper. The other is a much larger corpus of other EModE from a range of genres, the ESC: EEBO-TCP Segment, discussed fully in Murphy (2019).

As Murphy (2019) explains, a prototype approach was taken to the notion of genre in the ESLP. Lee (2001) and Taavitsainen (2001), amongst others, point out that terms such as 'genre' and 'text-type' are sometimes used freely and without clarification in research, leading to potential confusion. It is therefore important to be clear at the outset what is meant by such terms in any given piece of research. A distinction in perspective is made between the terms 'genre' and 'text-type' by Biber (1989: 15–16): if the categorisation of a text is determined by its external (situational) features, e.g. the function of the text and its audience(s), it is considered a genre; if by internal (linguistic) features, e.g. lexis and grammar, a text-type. The language in play-texts is shaped by its function: it is constructed for an audience and oriented to performance. These situational factors will have bearing on the investigation and analysis of the play-texts, so genre is the appropriate classification. In the ESLP more widely a genre is defined as "a category of texts grouped according to culturally, conventionally and consensually recognised criteria which change over time and which allow for division into sub-genres" (Murphy 2019: 62). In the prototype approach taken to the categorisation of ESLP texts, genres group into higher order 'domains', plays being part of the literary domain, along with poetry (Murphy 2019: 65). The plays genre includes sub-genres of comedy, tragedy and history (discussed further in Sections 2 and 5.3 in this article). Further categorisation of the types of plays within sub-genres, e.g. pastoral comedy or domestic comedy (see e.g. Mullan 2016), is not made formally in the ESLP corpora, but such play types are mentioned in my discussions of sub-genre (in Sections 2 and 5.3) as they can have bearing on language style. In this paper, 'style' refers to the language choices made by Shakespeare and other playwrights in their construction of characters, social groupings (people of different sex and/or social rank), dialogic and dramatic effects, plot and setting. For a wider discussion of terms commonly used to classify texts, including 'register', 'domain' and 'style' see Lee (2001).

In the rest of this article I begin by giving details of the content of the ESC: First Folio Plus in Section 2. Next, in Section 3 I explain why a new corpus was needed which is as closely relevant as possible to the content of the ESC: First Folio Plus. Here I point out that although several hundred extant play-texts from the Early Modern period are now freely available in digitised format, as one might expect they vary considerably in date, sub-genre, authorship and setting, all of which potentially have bearing on language style. In Section 4 I explain

briefly why other existing resources of EModE play data were not suitable for the ESLP. In Section 5 I give a brief outline of the steps involved in the compilation process (in 5.1), then I set out the inclusion criteria which were applied in order to select play-texts for the ESC: Comparative Plays. These are discussed in order of priority, beginning with date (in 5.2), then sub-genre (5.3), and finally other factors which were borne in mind alongside date and sub-genre to increase the relevance of the ESC: Comparative Plays content to the ESC: First Folio Plus content, and which constitute a set of less formal inclusion criteria (5.4). These other factors are authorship issues, audiences toward whom the plays were originally oriented, types of acting companies who first performed the plays, relative popularity, and verse/prose formatting. All are complex issues which, for reasons of space, can be discussed only briefly here. The final contents of the ESC: Comparative Plays are then set out in Section 6, followed by a brief reflection on the success of the compilation process in Section 7.

To enable the retrieval of quantitative, statistically-based results for the ESLP analyses, the ESC: First Folio Plus and the ESC: Comparative Plays needed to be incorporated into a customised Corpus Query Processor web-based corpus analysis interface (CQPweb), designed by Andrew Hardie, Lancaster University.[4] This required some essential formatting and re-formatting of the play-texts, in a series of post-processing stages explained in Section 8. The first of these details some fixes made to the downloaded, digitised play-texts (in 8.1). Secondly, in common with many corpus linguistic software tools, CQPweb relies on orthographic matching to generate accurate results, a process which is impeded by spelling variation that is typical in play-texts of this period. (English spelling was not fully standardised until later in the 17th century; see, for example, Nevalainen 2006: 32). Therefore, some standardisation of spelling in the corpora was necessary, discussed in 8.2. Finally, some mark-up and annotation of the play-texts was necessary to enable quantitative results to be retrieved through the CQPweb interface not only at word level, but also at grammatical and semantic levels, and to enable results to be restricted according to particular categories for comparison purposes (e.g. characters of different gender and/or social rank; plays of different sub-genre). This is detailed in 8.3. At the time of writing the tools which will generate results are in the final stages of development, and therefore no results from the corpora are included in this article; however, in due course see for example Archer and Gillings (in preparation), Culpeper and Findlay (in preparation) and Murphy et al. (in preparation).

2 The ESC: First Folio Plus

The ESLP's ESC: First Folio Plus comprises 38 plays: 36 from the First Folio (the first edited collection of Shakespeare's plays, published in 1623) plus *Pericles* and *The Two Noble Kinsmen,* which are now conventionally included as part of the Shakespeare 'canon' (in edited collections, such as *The Norton Shakespeare* (Greenblatt et al. 2016), and in critical discussions, such as Orlin 2003: 167–168).[5] The ESC: First Folio Plus was compiled from original-spelling electronic text files provided by Internet Shakespeare Editions (ISE).[6] The First Folio was chosen as the main source of Shakespeare's plays for the corpus because it was desirable to use a single source with reasonably consistent formatting and editing (both at the time of publication by printers and compositors, and more recently in the processes of transcription and digitisation). This was to make easier the reformatting of the texts to render them suitable for the CQPweb interface, and for the kinds of linguistic enquiry envisaged to meet the ESLP's aims. Reformatting texts with a relatively standard existing format means that more can be done using automated methods, and less by time-consuming manual adjustment (see 8.3).

Digitised versions of play-texts are based on editions which have been published, sometimes years after a play was first produced. A long gap between date of first production and publication potentially has bearing on language style change over time in the content of a play (or indeed any other text), and is therefore a consideration when creating a corpus designed to give a snapshot of language at a particular time. The notion of a 40-year period representing one 'generation' of language has been applied in the construction of some diachronic corpora, for example *A Corpus of English Dialogues 1560–1760* (the CED; see Kytö and Walker 2006) and the *Helsinki Corpus* (see Kytö 1996 [1991]). This notion was applied as a boundary to restrict the dates of texts in the synchronic corpora for the ESLP, to minimise the prospects of language change over time influencing the style of the content. The longest gap between date of first production and date of publication in the ESC: First Folio Plus is within one generation or 40-year span, at 33 years (*The Two Gentlemen of Verona*; see Table 1). Dates of first production are approximate in some cases, due to gaps in historical records. A further point regarding date is that in many cases multiple versions of a play exist, published in different years (and with likely editorial changes), so it is important to make clear which one is being used as the source for a corpus text. These matters are all relevant to ideas of when Shakespeare's plays can be said to have originated, and therefore also to the construction of a

comparative set of data that is designed to be 'contemporaneous' (discussed further in 5.2).

For the purposes of the ESLP the sub-genre designations of comedy, history and tragedy used in the First Folio of Shakespeare's plays are applied. In the ESC: First Folio Plus the 36 plays from the First Folio are assigned to the sub-genres designated therein, with the exception of *Cymbeline*, listed as a tragedy in the First Folio but reclassified as a comedy in the ESC: First Folio Plus for consistency with other plays which also feature comedy romance and a happy ending (*The Tempest*, for instance). *Pericles* and *The Two Noble Kinsmen* are designated as comedies, as is generally conventional (they are classified as comedies by Greenblatt et al. 2016, for instance). However, sub-genre classification is well known as not being straightforward, and the classification of some of Shakespeare's plays varies among scholarly sources and critical editions. As indicated in Section 1, play-texts in this period are sometimes classified into different types within each sub-genre (for example, pastoral comedy, city comedy and domestic comedy; revenge tragedy, classical tragedy and domestic tragedy). However, a formal classification of each play as one type or another is difficult to apply satisfactorily due to overlap in the features that typically define them (Mullan 2016 illustrates this in a discussion of the possible sub-types into which Shakespeare's comedies can be classified). Accordingly, the types of play in each sub-genre which Shakespeare particularly favoured (comedy with romantic themes and pastoral settings; historical dramas with British settings; revenge and classical tragedy) are borne in mind, but each play is not formally categorised according to type (see Orlin 2003 for more on the types of plays in each sub-genre of Shakespeare's works).

The play-texts in the ESC: First Folio Plus are listed in Tables 1 to 3, broken down by sub-genre, and listed in the order of the date they were first produced, beginning with the earliest. Dates of first production are from the *Database of Early English Playbooks* (DEEP), a useful scholarly and publicly accessible repository of online information about historical plays with search options for date, sub-genre, author and other variables.[7] Word counts (in number of tokens) are also given as a guide to the size of the body of data against which the ESC: Comparative Plays will be compared.[8]

Table 1: Comedy plays in the ESC: First Folio Plus

Play-text title	Date of first production (from DEEP)	Date of edition in corpus (from ISE)	Word count (from CQPweb)
Two Gentlemen of Verona	1590	1623	21212
The Taming of the Shrew	1592	1623	25344
The Comedy of Errors	1594	1623	17587
Love's Labour's Lost	1595	1623	25867
A Midsummer Night's Dream	1595	1623	20126
The Merchant of Venice	1596	1623	25065
The Merry Wives of Windsor	1597	1623	26663
Much Ado about Nothing	1598	1623	25203
As You Like It	1599	1623	25954
Twelfth Night	1601	1623	24033
All's Well that Ends Well	1603	1623	27423
Measure for Measure	1603	1623	26380
Pericles	1608	1609	22073
The Winter's Tale	1609	1623	31026
Cymbeline	1610	1623	33819
The Tempest	1611	1623	20482
The Two Noble Kinsmen	1613	1634	29393
Total word count			427650

Table 2: History plays in the ESC: First Folio Plus

Play-text title	Date of first production (from DEEP)	Date of edition in corpus (from ISE)	Word count (from CQPweb)
Henry VI, Part 2	1591	1623	30763
Henry VI, Part 3	1591	1623	29779
Henry VI, Part 1	1592	1623	26083
Richard III	1592	1623	35401
Richard II	1595	1623	26495
King John	1596	1623	24768

Henry IV, Part 1	1597	1623	29724
Henry IV, Part 2	1597	1623	31977
Henry V	1599	1623	31366
Henry VIII	1613	1623	30002
Total word count			296358

Table 3: Tragedy plays in the ESC: First Folio Plus

Play-text title	Date of first production (from DEEP)	Date of edition in corpus (from ISE)	Word count (from CQPweb)
Titus Andronicus	1592	1623	24584
Romeo and Juliet	1595	1623	29556
Julius Caesar	1599	1623	24037
Hamlet	1601	1623	34761
Troilus and Cressida	1602	1623	32060
Othello	1604	1623	32668
King Lear	1605	1623	29188
Timon of Athens	1605	1623	22510
Antony and Cleopatra	1606	1623	30277
Macbeth	1606	1623	21118
Coriolanus	1608	1623	33722
Total word count			314481

3 Why the new ESC: Comparative Plays was required

The ESLP uses a range of corpus linguistic techniques including the 'keyness' approach to obtain quantitative data for the analysis of language style. The keyness approach involves comparing one corpus to another on a statistical basis to identify language features (typically at the lexical, grammatical or semantic level) occurring with relatively high or low frequency (in this case in Shakespeare's plays) when compared to a reference corpus of other data (plays by a range of contemporaneous playwrights). A detailed discussion of keyness is outside the scope of this article (for this see, e.g., Baker 2004; Scott and Tribble 2006; Culpeper and Demmen 2015). However, research commenting on the content of reference corpora in keyness studies has bearing on the decision to undertake the (considerable) task of compiling the ESC: Comparative Plays.

Scott (2009) argues that a similar set of language features tends to be identified in a target corpus regardless of the content of the reference corpus. However, Culpeper (2009: 35), in his comparison of the dialogue of six characters in Shakespeare's play *Romeo and Juliet*, argues that "[t]he closer the relationship between the target corpus and the reference corpus, the more likely the resultant keywords will reflect something specific to the target corpus", a point that is also demonstrated in Fischer-Starcke's (2009) comparison of Jane Austen's novel *Pride and Prejudice* using three different reference corpora of varying closeness in related content. Fischer-Starcke shows that although some of the same results occur with all three reference corpora, the one with closely-relevant content helps identify some nuances that do not arise (statistically) with the others. Considering these studies, it seems likely that comparing Shakespeare's plays with another corpus of distantly-related content, even from the area of EModE drama, would yield results that are relatively general and afford little analytical potential (a problem envisaged by Crystal 2008: 21). A reference corpus comprised of closely-relevant content to the project's ESC: First Folio Plus improves the prospects for identifying features of Shakespeare's style that are or are not typical of playwriting of the period.

4 Existing sources of digitised EModE play-texts by other authors

As with the ESC: First Folio Plus texts obtained from ISE, it was desirable to obtain play-texts for the ESC: Comparative Plays from a single digitised resource with reasonably consistent formatting, to enable the necessary reformatting to be carried out as efficiently as possible (see 8.3). Early English Books Online (EEBO) was the resource used. In recent years, resources for comparative corpus-based research into EModE plays have increased substantially, notably through the digitisation of texts in a range of genres through the EEBO Text Creation Partnership (EEBO-TCP; 2000–2020). EEBO is publicly accessible, offers texts that are free to download, and is easily searchable through a number of portals offered by different providers. These tend to offer search fields such as: key word or term, date range, title, author(s), subject, and/or bibliographic reference numbers in the four collections from which the EEBO texts are sourced.[9] EEBO contains over 400 early extant play-texts from which to choose, amounting to approximately 3.25 million words of play-text data (estimated by Craig and Kinney 2009: xvii). The ProQuest LLC subscription database interface[10] was used for the ESC: Comparative Plays compilation because it offers searches on all the abovementioned fields, plus the benefit of facsimile manuscripts in the form of document images to which the corresponding digi-

tised text files are linked page by page. This made it easy to cross-reference between the two, which was crucial to the identification of missing or unclear text in the downloaded text files (discussed in 8.1).

Other repositories of EModE play-texts were considered, and rejected for a variety of reasons. Folger's *A Digital Anthology of Early Modern English Drama* (Brown et al. 2020) indexes the details of over 400 play-texts, but at the time of compilation only 29 were downloadable. More were required to compile a corpus of approximately the same size as the ESC: First Folio Plus (just over a million words; see Tables 1 to 3, Section 2). The EModE drama section of the *Vizualizing English Print* project[11] is larger than the Folger repository, offering 554 plays (including those from EEBO and more), but the spelling in the texts has been standardised in a manner beyond that required to meet the ESLP's goals. Original-spelling texts were required (as was the case for the ESC: First Folio Plus) so that spelling standardisation could be controlled and restricted to ensure that any language style features which might be of interest would be retained. Spelling adjustments are typically made to edited collections to render them more accessible or understandable, but can affect language style, especially if they are oriented to modernising the language (discussed further in 8.2). It is therefore an important consideration in choosing source material for any study of historical texts. The CED and Lutzky's (2012) EModE *Drama Corpus* are both oriented to diachronic linguistic research, and contain samples of play-text data. In contrast, the ESC: Comparative Plays required whole play-texts to match the ESC: First Folio Plus, i.e. with beginnings, middles and ends, entrances and exits to and from the stage, and other structural features which, if they do not occur in both corpora to a similar extent, may influence language style variation. Whole play-texts are included in the *Korpus of Early Modern Playtexts in English* (KEMPE; Petersen 2010: 164, 278–305) but, although publicly searchable, they are not downloadable and able to be manipulated.

5 Selection method and criteria for compiling the ESC: Comparative Plays

5.1 Steps in the corpus compilation process

First, a set of inclusion criteria was defined in consultation with the project's specialists in historical linguistics (Jonathan Culpeper, Lancaster University) and Renaissance drama (Alison Findlay, Lancaster University). As noted in Section 1 these are date, sub-genre and other less formal criteria (discussed in 5.2 to 5.4). Next, a list of extant EModE plays first produced between 1580 and 1625 was generated from DEEP (using the Advanced Search option). This returned

over 600 results, including multiple editions of many plays, with varying publication dates. The list was scrutinized and a shortlist of contender plays drawn up and ranked according to suitability of date, sub-genre and the other criteria. The EEBO interface was then searched using the date, title and author fields to determine which of the shortlisted plays were available in a digitised format. The earliest extant versions available (i.e. those with publication dates nearest the date of first production; see Section 2) were then downloaded and saved as text files. Full bibliographic details of the EEBO texts used in the ESC: Comparative Plays are given are given on the ESLP website.[12]

5.2 Date

Date was the most crucial criterion because the overarching requirement was for the ESC: Comparative Plays to be contemporaneous with the ESC: First Folio Plus. For the purposes of the ESLP, 'contemporaneous' means containing play-texts that originated during (approximately) the same historical period as those of Shakespeare, and within a period of 40 years (to represent a single generation of language, as explained in Section 2). Plays in the ESC: First Folio Plus were first produced between 1590 and 1613. For the search on DEEP the date parameters were increased by about a decade either side of the ESC: First Folio Plus play-texts, to 1580 to 1625, in anticipation of narrowing down a selection that was as close to a 40-year span as possible and which also fulfilled the sub-genre and other criteria. Only play-texts with a first production date and date of publication within 40 years of each other were considered. To increase parity with the spread of dates of play-texts in the ESC: First Folio Plus, efforts were made to include some that were first produced before 1600 and some first produced after 1600 in each sub-genre. Shakespeare's language style is argued as having changed around this time, both by linguists (e.g. Crystal 2008: 172) and by literary critics (e.g. Kermode 2000: 13, 45–46), and styles of other playwrights may well have changed over time (as indicated by Craig's 1999 study of Ben Jonson's plays).

5.3 Sub-genre

In addition to aiming for proportions of each sub-genre of similar size to those in the ESC: First Folio Plus, efforts were made to include a similar range of types of comedy, history and tragedy plays, especially pastoral and romantic comedy, revenge and classical tragedy, which feature strongly among Shakespeare's works. This was to reduce the prospects of language in the ESC: Comparative Plays being skewed towards language style(s) in particular types of play, a potential hazard relating to the kinds of characters typically encountered in par-

ticular sub-genres (e.g. royalty and nobility in history plays), their roles (e.g. shepherds in pastoral comedies), and/or the events which tend to occur (e.g. deaths in tragedies). Language variation of this kind is noted in existing corpus linguistic research into EModE play-texts by Hope (2010: 171), who states that "[c]ertain types of plot, and certain types of character, will entail certain types of vocabulary item – and there may even be syntactic expectations".

Pastoral comedies selected for inclusion in the ESC: Comparative Plays are, for example, Fletcher's *The Faithful Shepherdess* and Lyly's *Gallathea*. *Gallathea* has been compared to Shakespeare's *As You Like It* in other studies (e.g. Shapiro 2002: 318; Dillon 2003: 9). Domestic comedies include Wilkins' *The Miseries of Inforst Marriage* and Heywood's *How a Man May Chuse*. Romantic comedies include Heywood's *The Fair Maid of the West Part I*, the anonymous *Mucedorus*, Lyly's *Alexander and Campaspe*, plus *The Faithful Shepherdess* and *Gallathea* (mentioned above). A limited number of 'city' comedies are included (that is, plays with city settings and prototypical characters who inhabit them; see e.g. Braunmuller 2003: 60–61) because they usefully met the date criterion. The most prototypical are Jonson's *Bartholomew Fayre* and Middleton's *The Roaring Girl*. City comedies were popular at the time Shakespeare was writing, but not very typical of his own works. His play *The Merry Wives of Windsor* is considered as being near to a city comedy (Orlin 2003: 171–172), which gives some basis on which to include it in the ESC: Comparative Plays (although it is arguably less relevant than other types).

History plays are less variable in type than comedies and tragedies, although it is worth noting that while Shakespeare's histories all have British settings, those in the ESC: Comparative Plays include some with British settings (for example Marlowe's *Edward II* and Heywood's *Edward IV* Parts I and II) and some set elsewhere (e.g. *Tamburlaine Part I*). However, as comedy and tragedy plays in both corpora have a range of non-British and British settings, including history plays with non-British and British settings in the ESC: Comparative Plays is not inconsistent overall. Furthermore, most history plays with non-British settings included in the ESC: Comparative Plays allude to Britain's relationship with other countries at the time (e.g. Marlowe's *The Massacre at Paris* and Peele's *The Battle of Alcazar*) (see further Bartels 2003). They also include some British characters (Braunmuller 2003: 58–60), just as British-set plays sometimes involve characters from other countries. For example, Shakespeare's *Henry V* features dialogue in French and in French-accented English between Princess Katherine and her lady-in-waiting. The language variety used by characters seems more related to their sociolinguistic background than to the settings of the plays. Although there will be some lexical variation which corresponds to

different settings, language style at other levels (for example, grammatical or pragmatic) seems unlikely to be influenced much if at all.

Tragedies included in the ESC: Comparative Plays represent the main types found in the ESC: First Folio Plus: revenge tragedy (e.g. Kyd's *The Spanish Tragedy*; Webster's *The White Devil* and *The Duchess of Malfi*) and classical tragedy (Marlowe's *Dido, Queen of Carthage* and Jonson's *Sejanus*). A domestic tragedy is included (Heywood's *A Woman Killed With Kindness*), which fitted the date criterion, although this is another type which was restricted because of its limited presence in the ESC: First Folio Plus (*Othello* being the nearest; Orlin 2003: 171–172).

5.4 Other factors

As with the dating of plays and the classification of dramatic sub-genres, authorship is not entirely clear-cut and is frequently a contentious issue amongst critics. Potentially problematic issues for the ESC: Comparative Plays include certainty of authorship, collaborative authorship and author gender. Anonymously-authored plays were kept to a minimum, to reduce the future possibility of discovering overlapping authorship between the ESC: First Folio Plus and the ESC: Comparative Plays. However, it was necessary to include two anonymous plays to meet the date and sub-genre criteria. These are *The Life of Sir John Oldcastle* and *Mucedorus*. Shakespeare has been mooted as a possible author of *Mucedorus* (discussed in, for example, Wells et al. 1987 and Hope 1994), although without any firm supporting evidence to date. The authorship of a minority of other play-texts included has also been the subject of debate, such as *The Valiant Welshman* (attributed to Robert Armin).

Collaboration amongst playwrights in this period is known to be common, and even conventional (Thomson 2003: 49; Crystal and Crystal 2005: 57). Some of the plays included in the ESC: First Folio Plus are now widely accepted as having had input from other authors (e.g. *Henry VIII*, *Pericles* and *The Two Noble Kinsmen*), so collaborative plays which fitted the date and sub-genre criteria were also included in the ESC: Comparative Plays. These are Beaumont and Fletcher's *The Woman Hater*, *Philaster* and *The Maid's Tragedy*, and Middleton and Rowley's *The Changeling*.

Only works authored by male playwrights were included in the ESC: Comparative Plays. This was partly to avoid introducing the variable of author gender as a possible influence on language style (the ESC: First Folio Plus being entirely male-authored), but also because in this period women did not write or perform public drama (see e.g. Westfall 2002: 274; Braunmuller 2003: 62). Women did write for private performance, for example Lady Mary Wroth,

whose 1617 play *Love's Victory* was staged publicly for the first time in 2018.[13] However, unlike works for public performance, privately performed works were not subject to the approval of the Master of the Revels (a government official; see further Dutton 1991, 2000; Crystal and Crystal 2005: 62). It is therefore possible that language styles in plays written for private performance may be relatively less constrained than those in plays oriented to public performance. Shakespeare wrote for public theatres, although his plays and those of other male playwrights writing for public theatre were also sometimes performed at the royal court (Levin 2003: 101; Crystal and Crystal 2005: 7, 63, 181), for which some adjustments to the text may well have been made (for instance, to flatter the monarch). Dedications and prologues to be spoken in different performance settings are evident in some play-texts, but it is impossible to tell how much difference in character dialogue there would have been between public and court contexts.

In the period in which Shakespeare was writing, some acting companies were comprised of male adults and others were comprised of male children (female characters being played by younger men or boys; see e.g. Braunmuller 2003: 57–58). Plays first performed by companies of either age group are included in the ESC: Comparative Plays. There is no linguistic research to date which compares the language styles of plays first performed by children's companies and adult companies, but some literary critical research (Munro 2005: 2–3; Rutter 2012) suggests that the language content was not generally tailored or censored for younger performers. This seems reasonable given that some plays were performed by both adults' and children's companies.

As noted in Section 1, Shakespeare's plays were (and are) relatively popular and well known. For reasons of relevance, it was desirable for the ESC: Comparative Plays to represent works of similar esteem, although evaluating this is a somewhat subjective process. Works by a range of well-known contemporaneous playwrights are included (that is, who have been the focus of sustained critical interest and discussion), e.g. Ben Jonson, Thomas Kyd, Christopher Marlowe, John Marston and John Webster. Other playwrights who are arguably less well known are also represented, such as Thomas Drue, partly to fulfil the date and sub-genre criteria and partly to increase the diversity of authors represented overall. Diversity of authorship reduces the chance that language style in the ESC: Comparative Plays could be skewed by the over-representation of any particular authorial style features (such as those of Lyly's extravagant and mannered 'Euphuism' prose style; see e.g. Braunmuller 2003: 59–60). Author diversity was difficult to achieve, particularly in the tragedy section, because of the limited pool of digitised play-texts which met the date and sub-genre criteria.

Consequently, there is some over-representation of Marlowe's works relative to those of other playwrights.

Finally, efforts were made to include play-texts containing a mixture of verse and prose, to reflect the composition of Shakespeare's plays (see e.g. Thomson 2003: 47; Crystal and Crystal 2005: 165). However, it would have been impossible to match the proportions of verse and prose between the two corpora, given the constraints of other criteria, and the exact amounts of verse and prose have not been calculated. This should be borne in mind as a possible influence on results when the ESC: Comparative Plays is compared to the ESC: First Folio Plus (or indeed to any other dataset).

6 Contents of the ESC: Comparative Plays

The ESC: Comparative Plays comprises 46 play-texts by 24 authors: 20 comedies, listed in Table 4, 14 histories (Table 5) and 12 tragedies (Table 6). As for the ESC: First Folio Plus in Section 2, dates of first production and dates of the published version in the corpus can be compared side by side in the tables, and word counts (in number of tokens) are given.

Table 4: Comedy play-texts in the ESC: Comparative Plays

Author	Play-text title	Date of first production (from DEEP)	Date of edition in corpus (from EEBO)	Word count (from CQPweb)
John Lyly	*Alexander and Campaspe*	circa 1583	1584	15852
John Lyly	*Gallathea*	1585	1592	15992
Robert Greene	*Friar Bacon and Friar Bungay*	1589	1594	19142
George Peele	*The Old Wives Tale*	1590	1595	13175
George Chapman	*The Blind Beggar of Alexandria*	1596	1595	9662
Thomas Heywood	*The Fair Maid of the West Part I*	1604	1598	15200
George Chapman	*An Humerous Dayes Myrth*	1597	1631	19651
Henry Porter	*The Two Angry Women of Abington*	circa 1598	1599	31083
Anonymous	*Mucedorus*	1590	1599	30350
Thomas Dekker	*Old Fortunatas*	1599	1598	25767

Thomas Heywood	*How a Man May Chuse*	1602	1600	18820
Ben Jonson	*Volpone*	1606	1602	27501
Francis Beaumont and John Fletcher	*The Woman Hater*	1606	1616	29061
George Wilkins	*The Miseries of Inforst Marriage*	1606	1607	37449
Francis Beaumont	*The Knight of the Burning Pestle*	1607	1607	26140
John Fletcher	*The Faithful Shepherdess*	1608	1613	25878
Francis Beaumont and John Fletcher	*Philaster*	1609	1610	24411
Thomas Middleton	*The Roaring Girl*	1611	1620	30063
Ben Jonson	*Bartholomew Fayre*	1614	1611	46572
Philip Massinger	*The Bondman*	1623	1631	25389
Total word count				487158

Table 5: History play-texts in the ESC: Comparative Plays

Author	Play-text title	Date of first production (from DEEP)	Date of edition in corpus (from EEBO)	Word count (from CQP-web)
Robert Greene	*The Scottish History of James the Fourth*	circa 1590	1598	13180
Christopher Marlowe	*Tamburlaine Part I*	circa 1587	1590	21096
Christopher Marlowe	*Edward II*	1592	1594	24635
George Peele	*The Famous Chronicle of Edward I*	1591	1593	26347
Christopher Marlowe	*The Massacre at Paris*	1593	1594	25580
George Peele	*The Battle of Alcazar*	1589	1594	12352
Anthony Munday	*The Death of Robert Earl of Huntingdon*	1598	1601	28521
Thomas Heywood	*Edward IV Part I*	1599	1600	27723
Thomas Heywood	*Edward IV Part II*	1599	1600	29738

Anonymous	*The Life of Sir John Oldcastle*	1599	1600	26337
Thomas Heywood	*If You Know Not Me, You Know Nobody Part I*	1604	1605	13645
Thomas Dekker	*Sir Thomas Wyatt*	1602	1607	14731
Robert Armin	*The Valiant Welshman*	1612	1615	21048
Thomas Drue	*The Duchess of Suffolk*	1624	1631	20724
Total word count				305657

Table 6: Tragedy play-texts in the ESC: Comparative Plays

Author	Play-text title	Date of first production (from DEEP)	Date of edition in corpus (from EEBO)	Word count (from CQP-web)
Thomas Kyd	*The Spanish Tragedy*	1587	1592	19883
Christopher Marlowe	*The Jew of Malta*	1589	1633	16357
Christopher Marlowe	*Dr Faustus*	1592	1604	25539
Christopher Marlowe	*Dido, Queen of Carthage*	1586	1594	23071
Thomas Heywood	*A Woman Killed With Kindness*	1603	1607	14201
John Marston	*The Malcontent*	1604	1604	25057
Ben Jonson	*Sejanus*	circa 1604	1616	34544
Francis Beaumont and John Fletcher	*The Maid's Tragedy*	1610	1619	25431
John Webster	*The White Devil*	1612	1612	30846
John Webster	*The Duchess of Malfi*	1614	1623	30315
Thomas Middleton and William Rowley	*The Changeling*	1622	1653	30904
Thomas Middleton	*Women Beware Women*	1621	1657	22766
Total word count				298914

7 Reflections on the success of the compilation process

In general, the final content of the ESC: Comparative Plays fulfils the aim of creating a dataset that is contemporaneous with the ESC: First Folio Plus, and which meets the formal inclusion criteria devised to ensure its close relevance to the ESC: First Folio Plus (date and sub-genre, in consideration of other factors considered potentially to influence language style, discussed in Sections 5.2 to 5.4). Dates of first production (1584 to 1626) span a 42-year period beginning five years before the earliest play-text in the ESC: First Folio Plus (*The Two Gentlemen of Verona*, 1590) and ending thirteen years after the latest (*The Two Noble Kinsmen*, 1613). The gap between dates of production and publication for all play-texts in the ESC: Comparative Plays is within 40 years, and indeed most are within ten years. Therefore, the ESC: Comparative Plays can reasonably be considered to represent a single generation of language (discussed in Section 2), within which change over time has been minimised as a possible influence on language style. This is distinct from the possibility that the writing styles of the authors may have evolved over time, noted in 5.2. Efforts to include some plays in each sub-genre dated both pre- and post-1600, to match the ESC: First Folio Plus, were more successful in the comedy and history sections than in the tragedy section, where there were fewer digitised play-texts from which to choose that met the date criterion. Consequently, the tragedy section of the ESC: Comparative Plays is smaller than that of the ESC: First Folio Plus by about 15,500 words, whereas the comedy and history sections are larger, by just under 60,000 words and just under 10,000 words, respectively. For ease of reference, a side-by-side comparison of the size of the two corpora, broken down by sub-genre, is given in Table 7 (word counts again are from CQPweb).

Table 7: Side-by-side size comparison of the ESC: First Folio Plus and the ESC: Comparative Plays

	ESC: First Folio Plus	ESC: Comparative Plays
Comedy play-texts	427650	487158
History play-texts	296378	305657
Tragedy play-texts	314481	298914
All play-texts (total corpus size)	1038509	1091729

The relatively large size of the ESC: Comparative Plays comedy section partly reflects the fact that there were more digitised comedies to choose from which met the date criterion, so the deficit in the tragedy section could be offset by the inclusion of additional comedy data. The overall difference in size of the two

corpora of about 53,000 words can be accounted for by a combination of three factors. Firstly, unlike the Shakespeare play-texts which are nearly all from a single published edition (the First Folio), the ESC: Comparative Plays play-texts are nearly all from separate published editions. The latter more often feature extended and, in some cases, multiple prologues which precede the start of the first act/scene, and sometimes lists of dramatis personae at the end. The ESC: First Folio Plus play-texts feature much less of these kinds of text that are outside of the acts and scenes. During compilation, estimations of how much text to include in the ESC: Comparative Plays were based on dialogic text only, as this was anticipated as being the main source of comparison; word counts from CQPweb now include all text. Secondly, word counts vary between different software programmes, and during the compilation process these were made by a text editor, not by CQPweb (as now). Thirdly, word counts during the compilation process were based on original-spelling texts, and subsequent spelling normalisation has affected word count (e.g. through the closing up of open pronouns such as *him self*). The slight over-representation of comedy and under-representation of tragedy should be borne in mind in the interpretation of quantitative data, even if the difference in size of the datasets is automatically included in computations made by the corpus linguistic software tools (to assess whether any language style features more prevalent in comedies may be influencing results). Nevertheless, the sizes of each sub-genre section are sufficiently large to allow for sub-genre comparisons across the two corpora in the ESLP, and for future potential research (e.g. involving internal comparisons between sub-genre sections of the ESC: Comparative Plays).

While size is undoubtedly relevant to the comparability of the corpora and the sub-genre components, it is not the only basis on which comparability of content can be judged. An alternative perspective is given in Table 8, which compares the relative numbers of characters in each corpus (male, female and those whose gender did not clearly fit either category).

Table 8: Number of male and female characters in both corpora

	ESC: First Folio Plus				ESC: Comparative Plays			
	Comedy	History	Tragedy	All	Comedy	History	Tragedy	All
Male	445	440	380	1265	434	516	298	1248
Female	87	40	43	170	97	65	58	220
Unclassified	4	0	0	4	38	12	21	71
Total	501	480	458	1439	569	593	377	1539

As Table 8 shows, there is an overall difference of just 100 characters between the two corpora. This could simply be due to the ESC: Comparative Plays containing more plays (and therefore a greater diversity of characters) than the ESC: First Folio Plus, plays by Shakespeare being on average slightly longer than those by the other contemporaneous authors. Overall, the figures for male and female characters are quite similar across the corpora, though there were more characters in the ESC: Comparative Plays whose gender was difficult to classify. The figures in Table 8 only include characters' 'true' gender identities, not the assumed gender identities which some characters take on when disguised in the course of some of the plays (in both corpora). Where characters do assume alternative identities, it is marked up in the play-texts, because the language styles characters adopt when disguised may prove to be of interest for analysis.

8 Post-processing stages

8.1 Checks and fixes to the digitised play-texts

Some of the digitised play-text files downloaded from EEBO for the ESC: Comparative Plays contained missing words or parts of words, probably due to a lack of clarity in the printed manuscript versions from which they were transcribed. The extent to which this might impact on results would depend on how fine-grained an analysis was being made. However, to facilitate the most reliable results possible the play-texts for the ESC: Comparative Plays were checked by eye, and in many cases missing text was repaired (by examining the corresponding printed manuscript, as facsimiles on EEBO or in hard copy editions). There were also some textual anomalies consisting of the positioning of the ends of particularly long lines of text which, to conserve space in the printed manuscript, were sometimes placed by compositors at the right-hand margin in white space adjacent to the line immediately above or below the line to which they correspond. In some of the digitised EEBO texts (e.g. Webster's *The White Devil*) the bracketed line end is simply transcribed as part of the line next to which it appears, rather than inserted into the line in which it would actually be spoken. This was adjusted manually and the bracket removed, so that the running text of the plays is consistent with how it would be spoken. Whilst it departs from the formatting of the printed manuscript, it makes sense to do this for corpus texts in order to optimise the prospects for extracting results that rely on correct word order (e.g. multi-word units such as phrasal verbs). It also aids in the deployment of context-dependent automatic annotation tools (e.g. for adding grammatical part-of-speech tags; see 8.3).

8.2 Spelling normalisation

As mentioned in Section 1, spelling variation is common in texts of the period in which Shakespeare was writing, and some standardisation of word forms (usually referred to as spelling 'normalisation') is desirable because corpus linguistic software tools such as CQPweb rely on matching orthographic forms to generate results. For instance, the verb *would* has several different spelling variants (e.g. *would*, *woud*, *wud*) and if normalisation is not carried out the frequency counts for this verb would be split across all variants rather than aggregated, which could hamper accurate interpretation of its presence in texts.

The extent to which spelling normalisation is beneficial depends somewhat on text-type and research aims (further discussion of spelling normalisation in EModE corpora is given in Archer et al. 2015). For the ESLP it was important to standardise spelling to improve the prospects for orthographic matching as much as possible, but not to the extent that archaic words of potential interest would be lost. For example, the archaic verb form *holp* and the archaic plural noun form *eyne* were retained, not modernised respectively to *help* and *eyes*. The VARiant Detector spelling normalisation software (VARD 2; Baron and Rayson 2008) was used for the ESLP. Two researchers trained it on the ESC: First Folio Plus over a period of four months by first manually scrutinising every potential spelling variant in the Shakespeare play-texts, and then, using a set of guidelines (discussed in Demmen 2016), determining whether or not it should be adjusted. The trained version of VARD 2 was then deployed to normalise spelling variation automatically in the ESC: Comparative Plays, using a confidence threshold of 70 per cent (adjustable in the VARD 2 tool settings). This threshold was determined following tests with data samples, and was found to maximise the number of desirable spelling adjustments while minimising the number of undesirable adjustments (over-corrections or mis-corrections). Words for which VARD 2 recognises a high probability of being normalised in a certain way sometimes result in over-corrections. For example, the vast majority of cases of the highly frequent original spelling *bee* are instances of the verb *be*, and are correctly normalised accordingly. However, a minority are actually the noun *bee*, and these tend to get wrongly corrected to *be*, as in the following line from *Alexander and Campaspe* (Example 1):

(1) Be, he were best be as cunning as a <normalised orig="Bee" auto="false">Be</normalised>, or else shortly he will not be at all.

As Example (1) shows, the VARD 2 program uses tags to record spelling changes in the text (tagging is discussed in more detail in Section 8.3). The original spelling is preserved as the attribute <normalised orig> and the new, norma-

lised spelling is inserted before the end tag </normalised>. Similarly, the majority of cases of *deere* and *deare* are variant spellings of *dear*, but in the line from *Friar Bacon and Friar Bungay* in Example (2) *deere* should actually have been normalised to *deer*:

(2) The <normalised orig="mountaines" auto="true">mountains</normalised> full of fat and fallow <normalised orig="deere" auto="true">dear</normalised>,

Automated spelling normalisation using VARD 2 is less accurate than manual normalisation, but much faster. Since there was not sufficient time in the ESLP to normalise both the ESC: First Folio Plus and the ESC: Comparative Plays manually, automatic normalisation was used on the ESC: Comparative Plays. The relatively conservative confidence level used for the automated spelling normalisation of the ESC: Comparative Plays does mean that more words remain in original spelling than in the ESC: First Folio Plus. Despite some inaccuracies, it has undoubtedly improved the prospects for matching word forms across the two corpora.

8.3 Mark-up and annotation

As noted in Section 1, it was essential to mark up and annotate the play-texts of the ESC: First Folio Plus and the ESC: Comparative Plays. This was firstly to render the formatting suitable for incorporation into the CQPweb interface for the ESLP, and secondly so that the CQPweb corpus linguistic software tools could be used to select and restrict output according to variables such as subgenre, character gender and/or social rank category. Extensible Markup Language (XML; see Bray et al. 2008; Hardie 2014) tagging was used because it is a conventional technique for embedding metadata within a text (that is, information which can be used by other software programmes to read and interpret the text, and/or by researchers to record changes made to a text, e.g. through spelling normalisation). Also, XML tags can usually be excluded from display in results returned by corpus linguistic software tools, for convenience of viewing.

The mark-up of the corpora involved inserting XML tags to mark the start and end of each utterance in the dialogue of the play-texts, in the manner shown in Example (3) for the character Alsemero in Middleton and Rowley's *The Changeling*.

(3) <u who="CHANG_Alsemero" label="Als.">Not well indeed </u>

Utterance tags consist of an opening and a close 'u' tag (where 'u' stands for utterance). In the opening tag, the <u who> attribute is the character's name. It is

assigned by the researcher, remains consistent throughout the play-text, and is unique to a particular character. In contrast, the <label> attribute in the utterance start tag indicates the original speaker label in the play-text (if one is present). Speaker labels in EModE play-texts vary in consistency, often existing in several variant forms for a single character. Sometimes they are excluded, for example, if a speaker turn carries on after a stage direction. They are retained in the utterance tags through the <label> attribute but the more consistent <u who> attribute ensures that a single character's speech turns can all be captured by a search. The end tag </u> marks the utterance boundary. The 'u who' attributes refer to unique character identifiers which were indexed separately in a spreadsheet, to which other character attributes were linked (gender and social rank categories; the social rank classification system used was adapted from Archer and Culpeper (2003) and comprises eight categories: Monarch (rank 0), Nobility (1), Gentry (2), Professional (3), Other Middling Groups (4), Ordinary Commoners (5), Lowest Groups (6), Supernatural Beings (7), Problematic (8), discussed further in Murphy 2017). This metadata was not only essential to the analysis of the ESLP corpora; it also serves as a useful record of who's who in the play-texts for the researchers.

Stage directions, play titles, act and scene titles and boundaries (where present) were also marked up with XML tags, as were prologues and epilogues, bibliographic details, dedications, and any other material at the front or end of the play-texts such as lists of dramatis personae. This enables researchers to view the type(s) of text most pertinent to their enquiries without losing other textual details that might provide useful contextual information during the analysis process.

The insertion of XML tags was carried out by a team led by Andrew Hardie (Lancaster University). It was automated as far as possible (using language scripting tools and a text editor), where the format of items to be tagged was consistent, e.g. where existing speaker labels for a character were consistent (hence the choice of play-texts from single digital sources mentioned in Sections 2 and 4). Where the format was not consistent, e.g. for act and scene titles, tags were inserted manually.

As evident in Examples (1) and (2) in 8.2, XML tags were also inserted into the text by the VARD 2 spelling normalisation software, in both the manual and automatic modes, to annotate all spelling adjustments. A start tag captures the original spelling variant, and the standardised variant is inserted before the end tag, as shown in Example (4).

(4) <normalised orig="shal" auto="false">shall</normalised>

The ESC: First Folio Plus and the ESC: Comparative Plays were annotated with grammatical part-of-speech tags using the Constituent Likelihood Automatic Word-tagging System (CLAWS; see Leech et al. 1994).[14] CLAWS does not use XML tags, but alphanumerical codes in square brackets which correspond to over 200 part of speech classifications (in the CLAWS tagset version 6, used for the ESLP). For example, [JJ] denotes an adjective, [NN] a noun and [VV] a verb. The ESC: First Folio Plus play-texts were first tagged automatically using CLAWS, then scrutinized and corrected manually, then the CLAWS programme was adjusted (by Andrew Hardie, Lancaster University) so that it would interpret EModE more accurately (see further Demmen 2018). The adjusted CLAWS software was then deployed automatically over the ESC: Comparative Plays play-texts. As with the spelling normalisation procedure, this was because of limited time, and with the likelihood of some inaccuracy. However, it enables comparisons at the grammatical level across both corpora that would not otherwise be possible.

The two corpora have also been annotated with tags denoting the semantic meaning of each word (token), using the UCREL Semantic Analysis System (USAS; Rayson et al. 2004) in the Wmatrix suite of corpus linguistic software tools (Rayson 2008). USAS assigns a semantic category label in the form of an alphanumeric tag to each word, using a taxonomy of 232 categories of meaning grouped into 21 main semantic fields.[15] USAS has been successfully used for semantic analysis of historical texts (e.g. by Archer et al. 2009 and Culpeper 2011), despite being developed for late 20th century English. It has an option for EModE which utilises a lexicon extended to include common words in use in EModE such as *thou* and related forms. As with the spelling normalisation and grammatical tagging, the semantic tagging was done automatically and without post-correction due to time limitations.

9 Conclusion and future research prospects

In this article I have aired and discussed the issues involved in making comparisons between the language style(s) of Shakespeare's plays and those of plays by a range of other contemporaneous playwrights. I have done so from a corpus stylistic perspective, i.e. with a focus on the prospects for identifying language style features such as pragmatic or stylistic phenomena, rather than on authorship attribution in the computational stylistics tradition (which is the focus of much comparative corpus linguistic research into EModE plays to date, noted in Section 1). I have also taken in existing relevant research by linguists (e.g. Hope 1994, 2010; Crystal 2003, 2008) and by literary critical scholars (e.g. Dutton

1991, 2000; Orlin 2003), in light of the interdisciplinary nature of the ESLP and, potentially, other future research.

There are potential pitfalls in attempting to conduct comparative corpus-based research into language styles used by Shakespeare and other playwrights of his era, due to the diversity of date, sub-genre and other factors (discussed in Section 5) which could have bearing on results. Nevertheless, with careful management, the ESC: Comparative Plays has been compiled in ways which reduce the impact of those features as much as is feasible, to ensure it is as relevant as possible to the content of the ESC: First Folio Plus. This improves the prospects for identifying choices of language style features which Shakespeare did and/or did not share with other playwrights writing at around the same time.

I have illustrated some of the (many) difficulties of corpus compilation, particularly that which involves historical texts, and some of the inevitable compromises, notably the decision to offset a shortage of tragedy data that met the date criterion with some additional comedy data (mentioned in Section 7). I have also acknowledged some well-known difficulties with defining and determining dates of origination and sub-genre classifications for EModE plays (in Sections 2 and 5), and the thorny issues surrounding authorship (in 5.4). Uncertainty, anonymity and collaboration of authors inevitably blurs the lines a little between the ESC: First Folio Plus and the ESC: Comparative Plays. However, it is important to stress that distinguishing different authorial styles within plays is not the focus of the ESLP. Rather, it is the construction of different language styles of characters in two bodies of work. The ESC: First Folio Plus can be considered as one representing language substantially authored by Shakespeare and/or having a longstanding association with Shakespeare, whereas the ESC: Comparative Plays is one in which the language is substantially authored by others and not generally associated with Shakespeare.

Though the ESC: Comparative Plays was designed to meet the requirements of the ESLP, it is also well suited to other comparative corpus-based research where fine distinctions between Shakespeare's language style and the style of EModE plays more generally are sought. It facilitates quantitatively-based, comparative, corpus-based linguistic research in a range of areas which, as argued in Section 1, to date remain largely unexplored. Studies of pragmatic phenomena such as compliments, requests, forms of address and discourse markers could be carried out, as could comparisons of the types and uses of metaphor, discourses of nationhood and national identity and analyses of variation in language style between characters of different gender and social rank. Further insights into linguistic strategies for the construction of dramatic atmosphere and effects, such as suspense, might also be gained. Furthermore, the dearth of

records of natural speech surviving from the Early Modern period means that dramatic dialogue affords potential insight into wider spoken language of the time. Although dramatic dialogue differs from natural speech in some ways, e.g. through scripting and embellishment (Short 1996: 174–179), Culpeper and Kytö (2010: 17) argue that play-texts are nevertheless a "Speech-purposed" sub-genre "designed to produce real-time spoken interaction", and they demonstrate that spoken interaction can be productively investigated using historical play-text samples by a range of authors. Shakespeare's plays have also been the subject of sociolinguistic research by, for example, Brown and Gilman (1989) and Kopytko (1995). The ESC: Comparative Plays offers the potential for historical sociolinguistic research, and the possibilities are extensive and exciting. The ESC: Comparative Plays, together with the ESLP's ESC: First Folio Plus, will be publicly accessible, and therefore of potential benefit and use not only to academic researchers but to anyone else who may be interested, for example, school teachers, students, theatre groups and actors.

Notes

1. Dates of the 'Early Modern' period are considered to be circa 1500-1700 (Nevalainen 2006: 1).
2. 'Play-text' refers to the written form of a play under consideration (Culpeper and McIntyre 2006: 775), serving as a reminder that the focus is upon written, textual versions of plays, in contrast to performances.
3. See http://wp.lancs.ac.uk/shakespearelang/ (accessed 26.09.2019).
4. The CQPweb corpus analysis system is discussed further in Hardie (2012). See also https://cqpweb.lancs.ac.uk/ (accessed 26.09.2019).
5. These 38 plays are also listed by the Royal Shakespeare Company. See https://www.rsc.org.uk/shakespeares-plays/tragedies-comedies-histories (accessed 26.09.2019).
6. See http://internetshakespeare.uvic.ca/ (accessed 26.09.2019).
7. See http://deep.sas.upenn.edu/ (accessed 26.09.2019).
8. Word counts may be subject to slight change because, at the time of writing, the CQPweb interface and the corpora are not completely finalised.
9. The four EEBO source collections are *The English Short-Title Catalogue (1475–1640), compiled by A.W. Pollard and G.R. Redgrave (1927); the Short-Title Catalogue (1641–1700), compiled by Donald Wing (1945–1951); the Thomason Tracts (1640–1661) and the Early English Books Tract Supplement. See* http://eebo.chadwyck.com/about/about.htm *(accessed 26.09.2019).*

10. Accessible via subscription at https://search.proquest.com/eebo (accessed 26.09.2019).
11. *Visualizing English Print: Textual Analysis of the Printed Record*. 2016. http://graphics.cs.wisc.edu/WP/vep/ (accessed 26.09.2019).
12. http://wp.lancs.ac.uk/shakespearelang/files/2017/03/Comparative-corpus-reference-details.pdf (accessed 26.09.2019).
13. See http://wp.lancs.ac.uk/shakespeare-and-his-sisters/ (accessed 26.09.2019).
14. See http://ucrel.lancs.ac.uk/claws/ (accessed 26.09.2019).
15. See http://ucrel.lancs.ac.uk/usas/ (accessed 26.09.2019).

Acknowledgements

The research presented in this article was supported by the UK's Arts and Humanities Research Council (AHRC), grant reference AH/N002415/1. Before the end of 2020, the ESC: First Folio Plus and the ESC: Comparative Plays will be made publicly available, first via the CQPweb interface and then through download. I am grateful to Jonathan Culpeper, Alison Findlay, Sean Murphy and Brian Walker for their advice and comments.

References

Archer, Dawn and Derek Bousfield. 2010. 'See better, Lear'? See Lear better! A corpus-based pragma-stylistic investigation of Shakespeare's King Lear. In D. McIntyre and B. Busse (eds.). *Language and style*, 183–203. Basingtoke/New York: Palgrave Macmillan.

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, Jonathan Culpeper and Paul Rayson. 2009. Love – 'a familiar or a devil'? An exploration of key domains in Shakespeare's comedies and tragedies. In D. Archer (ed.). *What's in a word-list? Investigating word frequency and keyword extraction*, 137–157. Farnham/Burlington: Ashgate.

Archer, Dawn and Mathew Gillings. In preparation. Depictions of deception, focussing on five Shakespearean characters.

Archer, Dawn, Merja Kytö, Alistair Baron and Paul Rayson. 2015. Guidelines for normalising Early Modern English corpora: Decisions and justifications. *ICAME Journal* 39 (1): 5–24.

Baker, Paul. 2004. Querying keywords. Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics* 32 (4): 346–359.

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. http://ucrel.lancs.ac.uk/vard/about/ (accessed 26.09.2019).

Bartels, Emily. 2003. Shakespeare's view of the world. In S. Wells and L. Cowen Orlin (eds.). *Shakespeare. An Oxford guide*, 151–164. Oxford: Oxford University Press.

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27 (1): 3–43.

Braunmuller, Albert. 2003. Shakespeare's fellow dramatists. In S. Wells and L. Cowen Orlin (eds.). *Shakespeare. An Oxford guide*, 55–66. Oxford: Oxford University Press.

Bray, Tim, Jean Paoli, Michael 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/ (accessed 26.09.2019).

Brown, Meaghan, Michael Poston and Elizabeth Williamson (eds.). 2020. *A digital anthology of Early Modern English drama*. Folger Shakespeare Library. http://emed.folger.edu (accessed 26.09.2019).

Brown, Roger and Albert Gilman. 1989. Politeness theory and Shakespeare's four major tragedies. *Language in Society* 18: 159–212.

Busse, Beatrix. 2006. *Vocative constructions in the language of Shakespeare.* Amsterdam/Philadelphia: John Benjamins.

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

CED = *A Corpus of English Dialogues 1560–1760*. 2006. Compiled by Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University).

Craig, Hugh. 1999. Jonsonian chronology and the styles of *A Tale of a Tub*. In M. Butler (ed.). *Presenting Ben Jonson: Text, history, performance*, 210–232. Basingstoke: Palgrave Macmillan U.K.

Craig, Hugh and Arthur Kinney (eds.). 2009. *Shakespeare, computers, and the mystery of authorship.* Cambridge: Cambridge University Press.

Crystal, David. 2003. The language of Shakespeare. In S. Wells and L. Cowen Orlin (eds.). *Shakespeare. An Oxford guide*, 67–78. Oxford: Oxford University Press.

Crystal, David. 2008. *Think on my words. Exploring Shakespeare's language.* Cambridge: Cambridge University Press.

Crystal, David and Ben Crystal. 2005. *The Shakespeare miscellany.* London: Penguin.

Culpeper, Jonathan. 2009. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics* 14 (1): 29–59.

Culpeper, Jonathan. 2011. A new kind of dictionary for Shakespeare's plays: An immodest proposal. In M. Ravassat and J. Culpeper (eds.). *Stylistics and Shakespeare's language. Transdisciplinary approaches*, 58–83. London/ New York: Continuum.

Culpeper, Jonathan and Jane Demmen. 2015. Keywords. In D. Biber and R. Reppen (eds.). *The Cambridge handbook of English corpus linguistics*, 90–105. Cambridge: Cambridge University Press.

Culpeper, Jonathan and Alison Findlay. In preparation. Contemporary understandings of Welsh, Scottish and Irish identities: Celtic characters in Shakespeare's *Henry V.*

Culpeper, Jonathan and Merja Kytö. 2010. *Early Modern English dialogues: Spoken interaction as writing.* Cambridge: Cambridge University Press.

Culpeper, Jonathan and Dan McIntyre. 2006. Drama: Stylistic aspects. In K. Brown (ed.). *Encyclopedia of language and linguistics.* Volume 3, 772–785. 2nd edition. Oxford: Elsevier.

Demmen, Jane. 2016. Smoothing out spelling variation. Blog post 22.10.2016. http://wp.lancs.ac.uk/shakespearelang/2016/10/22/smoothin-out-spelling-variation/ (accessed 26.09.2019).

Demmen, Jane. 2018. Is that a verb I see before me? Implementing grammatical category/part-of-speech tagging in the Shakespeare Corpus. Blog post 20.06.2018. http://wp.lancs.ac.uk/shakespearelang/2018/06/20/is-that-a-verb-i-see-before-me-implementing-grammatical-category-part-of-speech-tagging-in-the-shakespeare-corpus/ (accessed 26.09.2019).

Dillon, Janette. 2003. Shakespeare and the traditions of English stage comedy. In R. Dutton and J.E. Howard (eds.). *A companion to Shakespeare's works.* Volume III. *The comedies*, 4–22. Malden/Oxford/Victoria: Blackwell.

Dutton, Richard. 1991. *Mastering the revels*. Basingstoke/London: Macmillan.

Dutton, Richard. 2000. *Licensing, censorship and authorship in Early Modern England*. Basingstoke/New York: Palgrave.

EEBO-TCP = *Early English Books Online-Text Creation Partnership*. 2020. https://www.textcreationpartnership.org/ (accessed 26.09.2019).

Fischer-Starcke, Bettina. 2009. Keywords and frequent phrases of Jane Austen's *Pride and Prejudice*. A corpus-stylistic analysis. *International Journal of Corpus Linguistics* 14 (4): 492–523.

Greenblatt, Stephen, Walter Cohen, Suzanne Gossett, Jean Howard, Katherine Eisaman Maus and Gordon McMullan (eds.). 2016. *The Norton Shakespeare*. 3rd Edition. London/New York: W.W. Norton and Company.

Hardie, Andrew. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17 (3): 380–409.

Hardie, Andrew. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal* 38: 73–103.

Helsinki Corpus = The Helsinki Corpus of English texts. 1991. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka and Matti Kilpiö (Old English); Saara Nevanlinna and Irma Taavitsainen (Middle English); Terttu Nevalainen and Helena Raumolin-Brunberg (Early Modern English). Helsinki: Department of English, University of Helsinki.

Hope, Jonathan. 1994. *The authorship of Shakespeare's plays*. Cambridge: Cambridge University Press.

Hope, Jonathan. 2010. *Shakespeare and language: Reason, eloquence and artifice in the Renaissance*. London: Arden Shakespeare.

Hope, Jonathan and Michael Witmore. 2010. The hundredth psalm to the tune of 'Green Sleeves': Digital approaches to the language of genre. *Shakespeare Quarterly* 61 (3): 357–390.

KEMPE = *Korpus of Early Modern Playtexts in English*. Initially compiled by Lene B. Petersen and Marcus X. Dahl, in association with Visual Interactive Syntax Learning (VISL), Southern Denmark University (SDU), 2001–2003. The fully searchable version of the corpus was prepared by Lene B. Petersen and Eckhard Bick, July 2004.

Kermode, Frank. 2000. *Shakespeare's language*. London: Penguin.

Kopytko, Roman. 1995. Linguistic politeness in Shakespeare's plays. In A.H. Jucker (ed.). *Historical pragmatics. Pragmatic developments in the history of English*, 515–540. Amsterdam/Philadelphia: John Benjamins.

Kytö, Merja. 1996 [1991]. *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts.* 3rd edition. Helsinki: Department of English, University of Helsinki.

Kytö, Merja and Terry Walker. 2006. *Guide to* A Corpus of English Dialogues 1560–1760 (Studia Anglistica Upsaliensia 130). Uppsala: Acta Universitatis Upsaliensis.

Lee, David. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3): 37–72.

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 (accessed 26.09.2019).

Levin, Carole. 2003. The society of Shakespeare's England. In S. Wells and L. Cowen Orlin (eds.). *Shakespeare. An Oxford guide*, 93–102. Oxford: Oxford University Press.

Lutzky, Ursula. 2012. *Discourse markers in Early Modern English*. Amsterdam/Philadelphia: John Benjamins.

Mullan, John. 2016. An introduction to Shakespeare's comedy. British Library online article published 15.03.2016. https://www.bl.uk/shakespeare/articles/an-introduction-to-shakespeares-comedy (accessed 26.09.2019).

Munro, Lucy. 2005. *Children of the Queen's Revels. A Jacobean theatre repertory*. Cambridge: Cambridge University Press.

Murphy, Sean. 2017. Shakespeare and social status. Blog post 05.06.2017. http://wp.lancs.ac.uk/shakespearelang/2017/06/05/shakespeare-and-social-status/ (accessed 26.09.2019).

Murphy, Sean. 2019. Shakespeare and his contemporaries: Designing a genre classification scheme for *Early English Books Online* 1560–1640. *ICAME Journal* 19: 59–82.

Murphy, Sean, Jane Demmen, Alison Findlay and Dawn Archer. In preparation. Mapping the links between gender, status and genre in Shakespeare's plays.

Nevalainen, Terttu. 2006. *An introduction to Early Modern English*. Edinburgh: Edinburgh University Press.

Orlin, Lena Cowen. 2003. Part II Shakespearean genres: Introduction. In S. Wells and L. Cowen Orlin (eds.). *Shakespeare. An Oxford guide*, 167–174. Oxford: Oxford University Press.

Petersen, Lena. 2010. *Shakespeare's errant texts. Textual form and linguistic style in Shakespearean 'bad' quartos and co-authored plays.* Cambridge: Cambridge University Press.

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), 25 May 2004, Lisbon, Portugal, 7–12. Paris: European Language Resources Association.

Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13 (4): 519–549.

Rutter, Carol Chillington. 2012. Playing with boys on Middleton's stage – and ours. In G. Taylor and T.T. Henley (eds.). *The Oxford handbook of Thomas Middleton*, 98–115. Oxford: Oxford University Press.

Scott, Mike. 2009. In search of a bad reference corpus. In D. Archer (ed.). *What's in a word-list? Investigating word frequency and keyword extraction*, 79–91. Oxford: Ashgate.

Scott, Mike and Chris Tribble. 2006. *Textual patterns*. *Key words and corpus analysis in language education.* Amsterdam/Philadelphia: John Benjamins.

Shapiro, Michael. 2002. Boy companies and private theatres. In A.F. Kinney (ed.). *A companion to Renaissance drama*, 314–325. Oxford/Malden: Blackwell.

Short, Mick. 1996. *Exploring the language of poems, plays and prose.* London/ New York: Longman.

Taavitsainen, Irma. 2001. Changing conventions of writing: The dynamics of genres, text types, and text traditions. *European Journal of English Studies* 5(2): 139–150.

Thomson, Peter. 2003. Conventions of playwriting. In S. Wells and L. Cowen Orlin (eds.). *Shakespeare. An Oxford guide*, 44–54. Oxford: Oxford University Press.

Wells, Stanley, Gary Taylor, John Jowett and William Montgomery. 1987. *William Shakespeare: A textual companion.* Oxford: Clarendon Press.

Westfall, Suzanne. 2002. "What revels are in hand?" Performances in the great households. In A.F. Kinney (ed.). *A companion to Renaissance drama*, 266–280. Oxford/Malden: Blackwell.