A corpus stylistic investigation of the language style of Shakespeare's plays in the context of other contemporaneous plays

Jane Elizabeth Judson Demmen, BA Hons, MA

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Linguistics and English Language
Lancaster University, U.K.
September 2012

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

A corpus stylistic investigation of the language style of Shakespeare's plays in the context of other contemporaneous plays

Jane Elizabeth Judson Demmen, BA Hons, MA

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Dept. of Linguistics and English Language, Lancaster University, U.K. September 2012

ABSTRACT

Shakespeare's plays occupy a uniquely prominent position in English language and literature. Shakespeare was, however, one among a number of other successful and popular playwrights of the late 16th and early 17th centuries, and, when examined on an empirical basis, his language style has much in common with that of his peers. In this corpus stylistic study, I investigate similarities and differences between the language in Shakespeare's plays and in a range of plays by a selection of other contemporaneous dramatists. My quantitative data is extracted from an existing corpus containing Shakespeare's First Folio, and a new, specialised parallel corpus of plays from similar dates and genres written by other contemporaneous dramatists. This new corpus was constructed during the study.

The corpus linguistic methods I use are simple frequency, keyness (Scott e.g. 1999, 2000) and Baker's (2011) new concept of "lockwords". Simple frequency and keyness (linguistic items occurring with comparatively low or high statistical frequency) are established corpus linguistic methods for investigating language styles in literary texts. However, as Baker (2004:349) argues, keywords highlight only the differences between texts. Similarities are also important, to contextualise differences and avoid overstating

their stylistic implications. Moreover, as I show in this study, empirical evidence of similarities is of stylistic interest. It reveals preferences for language style features which Shakespeare and other contemporaneous dramatists shared, and which constitute features of the register of Early Modern English drama. I examine three types of language units in each corpus: single words, word clusters and semantic domains. I extract word and word cluster data using Scott's (1999) *WordSmith Tools* and semantic domain data using Rayson's (2009) *Wmatrix* software tools.

My findings have implications for (a) the distinctiveness of Shakespeare's style, (b) the register of EModE drama and (c) methods for investigating language similarities using corpus linguistic methodology.

DECLARATION

I declare that this is my own work, and that it has not been submitted in substantially the same form for the award of a higher degree elsewhere.

Jane Elizabeth Judson Demmen, BA Hons, MA

Department of Linguistics and English Language
Lancaster University, U.K.
September 2012

ACKNOWLEDGEMENTS

I am very grateful for the award of an Arts and Humanities Research Council doctoral studentship (award number 2009/ PG144956).

I owe much thanks to many people at Lancaster University who have contributed to my research. I am especially grateful to my supervisor Jonathan Culpeper, who has been so generous with his time, help and advice. I would also like to thank Alistair Baron for tuition on the *VARD 2* software, Andrew Hardie for his expertise with PHP, and Paul Baker, Paul Rayson and Andrew Wilson for their guidance over corpus methods and statistical matters. I am also grateful to Alison Findlay and Liz Oakley-Brown for their expert insight into Renaissance drama. I thank Marjorie Wood for managing the many administrative matters relating to my PhD with great kindness and efficiency.

I would like to thank Mike Scott for providing access to his Shakespeare corpus and for occasional specialist advice about *WordSmith Tools*.

I have been buoyed up during my studies by my family, academic colleagues and friends. I would particularly like to thank Brian Walker for his unflagging interest in the project over the last three years, for sharing his expertise on *Wmatrix*, and for numerous useful discussions which have helped me to understand my research more clearly. My draft work has also greatly benefited from his critical comments and suggestions. I am very grateful to Shaun Austin, Sue Burling, Ursula Lutzky and Paul Rayson for reading parts of my thesis, which is much improved as a result of their feedback.

Finally, I could not have undertaken the PhD without Phillip Demmen, "the loyal'st husband that did e'er plight troth" (Shakespeare, *Cymbeline*, I:i). He has supported me throughout my studies in countless ways, for which I am truly grateful.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 The significance of this study

Shakespeare's plays occupy a unique position in English language and literature. In addition to their longstanding popularity in performance, they have been the subject of continued scholarly interest and debate, particularly in the literary critical tradition. The development of the field known as "digital humanities" in the late 20th and early 21st centuries has facilitated the investigation of Shakespeare's plays in new ways, on a statistical basis, via the rapid, automated and systematic quantitative analysis of language features. In recent years, linguists have begun to research the plays using corpus linguistic software tools which deploy complex algorithms to identify, count, compare and categorise language features in electronic files of digitised texts. Scholars who have applied corpus methods to the linguistic investigation of Shakespeare's plays include, for example: Archer and Bousfield (2010); Archer et al. (2009); Craig (2004, 2010, 2011); Craig and Kinney and other contributors to their (2009) edited volume; Culpeper (2002, 2009, 2011); Hope and Witmore (2004, 2010); Petersen (2010); Scott and Tribble (2006:59-70). Their research ranges from close comparisons of a selection of characters in a single play (e.g. Archer and Bousfield 2010; Culpeper 2002, 2009) to analyses of language features based on all of Shakespeare's plays (e.g. Hope and Witmore 2010; Scott and Tribble 2006).

Shakespeare was, however, one among a number of successful and popular playwrights writing in the late 16th and early 17th centuries (Crystal and Crystal 2005:142). Yet linguistic research which examines the language of his plays in the context of other drama of the period on an empirical basis is scarce, as pointed out by Craig (2011:53) in his computational stylistic study of the vocabulary of Shakespeare and other contemporaneous dramatists. Hope and Witmore (2010:387-390), who map

the relative similarity of rhetorical features in 320 plays from the Early Modern period, including those of Shakespeare, also argue that there is a need for much further comparative work (2010:387). Culpeper (2011) concurs, concluding from his initial corpus stylistic investigations of Shakespeare's language style in the context of other playwrights of the period that a much bigger study is called for (which could culminate in a comparative, corpus-based dictionary for Shakespeare's plays).

The abovementioned studies, and the small body of other existing corpus research which compares Shakespearean[1] and other contemporaneous drama on a statistical basis, have begun to put the vast bank of existing literary critical research into Early Modern English ("EModE") drama into some empirically-based perspective. For example, Craig (2011, 2012), Crystal (2008), Elliott and Valenza (2011) and Rosso et al. (2009) argue that quantitative corpus data shows the vocabulary of Shakespeare's plays to be similar to that of other drama of the period. According to Craig (2011:53-58) and Crystal (2008:2-6), this evidence counters the longstanding and popular idea that Shakespeare had an exceptionally large and/or inventive vocabulary (suggested by, for example, Greenblatt 1997:63 and Marche 2011:35). In acknowledging Shakespeare's undoubted skill in using language, Craig (2012:4) states that "Shakespeare does have a distinctive style [...] but there is no evidence that Shakespeare is somehow more distinctive than anyone else". Crystal (2008:232-233) argues that the distinctiveness of Shakespeare's language lies in its "effective bending and breaking of rules", and that "[e]conomy of expression, the result always of a trading relationship between lexicon and grammar, is the hallmark of Shakespeare's linguistic creativity". From these findings, there is clearly much more to be learned about Shakespeare's language style by comparing it with that of his

[1] In this study, "Shakespearean" drama means plays authored by Shakespeare (not pertaining to or in the manner of Shakespeare, which are possible meanings in other contexts).

contemporaries on an empirical basis, and particularly by going beyond the lexical level. This is where my study makes a contribution, using a corpus stylistic approach that includes some new and some more established corpus linguistic methods (explained fully in chapters 2 and 3).

To carry out my corpus stylistic study of Shakespeare's language style, I use an existing corpus of the plays in Shakespeare's First Folio, the texts of which are already annotated so as to be suitable for corpus linguistic methods (see 4.2.2 and 5.2). This is an adapted version of Mike Scott's Shakespeare corpus[2] which was prepared for my previous (2009) research, which I refer to as the *Shakespearean Drama Corpus*[3] (also abbreviated to "*SDC*"). I discuss its background and suitability for the present study in 4.2. Using it allowed me the time and space in the study to construct a new, specialised "parallel" corpus of other contemporaneous plays for comparison with Shakespeare's First Folio, using digitised texts from *Early English Books Online* (hereafter "*EEBO*") (see 4.3.1). I follow Leech and Smith (2005) in using the term "parallel" to mean corpora which are closely comparable in content[4].

In order to see what kinds of language features are distinctive in Shakespeare's plays, not just in EModE plays in general, it is important that the Shakespeare corpus is compared to a "reference corpus" of very closely-related content (Culpeper 2009:35). The parallel corpus I constructed for this study contains EModE dramatic dialogue of similar date and genre to Shakespeare's plays, in order to maximise the relevance of the results. I discuss this fully in chapter 4, where I argue that no existing corpus of EModE plays is sufficiently closely comparable to be suitable for my study.

---

[2] See http://www.lexically.net/wordsmith (accessed 05.08.12).
[3] "Drama" is limited to plays in this study, though in wider contexts it also encompasses other types of entertainment (masques, for example).
[4] Leech and Smith (2005:95) note that the term "parallel" is often used to describe corpora containing translations of the same texts in more than one language, for which they argue the terms "matching" or "equivalent" are more appropriate.

I refer to the new corpus as the *Non-Shakespearean Early Modern English Drama Corpus* (or "*NDC*").

The data on which I base my analyses are extracted from the corpora as three types of language units, using three statistical methods. The language units are:

(i) single words (as these are identified by the corpus linguistic software tools: essentially strings of letters within boundaries of spaces and/or punctuation);

(ii) word clusters (electronically-derived recurrent word combinations based on collocational relationships, i.e. words which frequently occur near one another; see Scott 1999:Help menu and my further discussions in 3.2); and

(iii) semantic domains (groups of words which are related semantically, again as identified and classified by the corpus tools, discussed further in 3.3).

The methods I use are discussed fully in 2.5 and 2.6. These are:

(i) frequency lists: language units which occur with the highest frequency (when the corpora are analysed independently of one another);

(ii) keyness: language units which are "key" in Shakespeare's plays when they are compared to the other contemporaneous plays, i.e. those which occur with statistically high or low frequency (Baker 2004; Scott e.g. 1999, 2000); and

(iii) locking: language units which occur with the most similar high frequency, statistically (building on Baker's 2011 new concept of "lockwords").

I follow other corpus stylisticians in using the corpus linguistic software *WordSmith Tools* (Scott 1999) (hereafter "*Wordsmith*") to obtain single-word and word cluster data, and *Wmatrix* (Rayson 2009) ("*Wmatrix*") to extract semantic domain data.

The decision to include high-frequency and key results is based on evidence from existing corpus stylistic studies that the output points to potential style features which reward closer qualitative analysis (e.g. Archer et al. 2009; Culpeper 2002, 2009,

2011; Ho 2011; Mahlberg 2007 and McIntyre 2010), as I discuss fully in chapters 2 and 3. Word clusters and semantic domains are still relatively under-researched in EModE plays. The decision to use them in this study is informed by existing research that shows they are useful for the stylistic analysis of literary texts. It is also supported by the arguments of David Crystal (from his extensive corpus-based research) that:

(i) collocations merit more attention in Shakespeare's plays (2008:173); and

(ii) grouping words in Shakespeare's plays according to semantic meaning allows for a clearer view of concepts which may be unfamiliar in the present day (2008:155).

I include Baker's (2011) new "locking" method in my study so that I can examine similarities in the language style of Shakespeare and his peers, as well as the differences which are highlighted by key results. In his discussion of keyness analyses, Baker (2004:349) points out that key results provide no information about similarities between texts, only about differences, and that similarities should not be ignored. Similarities provide a context against which differences in language styles can be seen. This gives a more balanced perspective, as there is a risk of overstating language differences if similarities are not also brought into the frame. Similarities also have implications for theories of "foregrounding" (Jakobson 1960; Mukařovský 1964a and b; see also van Peer 1986), which have become a cornerstone of stylistic analysis. Foregrounded language is argued as being that which stands out as noticeable to a reader (psychological foregrounding, in Leech and Short's 2007 terminology), or by extension to a listener or audience, by virtue of deviating from language norms. This creates a "defamiliarization" effect (see e.g. van Peer 1986:3-4). Foregrounded language can be identified through quantitative analysis of language differences, by using the keyness technique to find linguistic items which deviate, statistically, from

norms in a particular text or set of texts (see e.g. Mahlberg and McIntyre 2010:207). Keyness analysis provides no information about the language which could be said to represent the norms of a text or texts from which these items deviate, however. An examination of similarities using a technique such as the "locking" method is therefore important, in order to provide some statistically-based information about the **non-**deviant language, which arguably constitutes the norms against which foregrounded language can be seen. I write "arguably" because there is no guarantee that statistical similarities between linguistic items in one body of text and another constitute a relevant cognitive background for what is psychologically foregrounded (see 7.5.3).

With the exception of Ho (2011), there are to date no corpus stylistic studies that apply statistical methods to investigate similarities alongside differences in the language in literary texts. This study begins to address the shortfall in research into similarities in language styles. I explain in 2.6 that Baker's new concept of "lockwords" enables an investigation of similarities on a statistical basis in my study, because lockwords are "the opposite of Scott's (2000) concept of keywords" (Baker 2011:73). My study extends Baker's (2011) research by:

(i) applying the locking principle to other types of language constructs apart from words (to word clusters and semantic domains);

(ii) using it with synchronic corpora; and

(iii) testing it with historical texts.

Since I have only two corpora, I adapt the keyness tools of *WordSmith* and *Wmatrix* to apply the locking method, rather than using Baker's (2011) methods of standard deviation and co-efficient of variance (to investigate four diachronic corpora).

Above, I have argued that my study helps address the scarcity of corpus stylistic research which examines Shakespeare's language style in the context of that of

his peers. It takes the initial comparative research carried out by Culpeper (2011) much further. It also adds to what has been found in the computational stylistic area, which is oriented more to authorship attribution and less to the qualitative analysis of pragmatic and discoursal effects of language features highlighted by quantitative data (discussed further in 2.4.3). My study tests out a new method which complements keyness by focusing on similarities (the locking concept), to see what benefit it offers to corpus stylistic investigations. It also adds to the relatively small amount of existing research into EModE plays which goes beyond single-word language units, by investigating word clusters and semantic domains. Finally, the new, specialised parallel reference corpus for Shakespeare's First Folio will be a useful resource for future research as well as the present study.

In the next section, I set out the aims of the study in more detail. I also make clear the scope of my research, given that the possibilities for exploring the language style of Shakespearean and other EModE plays are enormous, but that time and writing space are limited.

1.2 Aims and scope of the study

As outlined in the previous section, this is a corpus stylistic study with the following overarching aims:

(i) to build a parallel reference corpus for Shakespeare's plays;

(ii) to begin exploiting the new corpus using a variety of statistical methods, in order to identify and explore similarities and differences in the style of language in plays created by Shakespeare in comparison to a range of his contemporaries; and

(iii) to draw some conclusions about the value of the different methods applied, especially the new concept of statistical "locking" (Baker 2011) in order to see what it adds to an investigation of simple frequencies and key results.

These aims form the nuclei of my research questions, given in the next section.

Although my study has implications for Shakespeare's individual authorial style, in the context of a group of other contemporaneous playwrights, it does not extend to the attribution of authorship (in the manner of computational stylistic research such as Craig and Kinney 2009 and Petersen 2010, mentioned in the previous section). My findings are oriented to explaining the meaning and effect of language in the way that it is used by Shakespeare and other playwrights of his day, and the extent to which they share style preferences that appear to be characteristic of the register of EModE drama (rather than to asserting that dramatic dialogue did or did not originate from particular authors).

Crystal (2008) is of the view that it is preferable to focus on internal variation in Shakespeare's plays (e.g. by contrasting characters of different gender, social rank, genre and/or in earlier or later stages of Shakespeare's writing career), instead of making comparisons with other corpora. He argues that:

> Given the extraordinary range of character and content in Shakespeare, and the period of time (over twenty years) over which he wrote, valid stylistic generalizations are likely to be impossible – or, at least, to be of such generality to be uninformative (2008:21).

Craig's (1999) research into Ben Jonson's plays also shows that an author's style does not remain static. How then can any overall comparisons of style be made between Shakespeare and his peers? A comparison of Shakespeare's plays with a corpus of plays with very different content and dates, even from within the register of EModE drama, would be likely to yield the "uninformative" kinds of results which Crystal mentions, since reasons to do with dating, genre, themes and intended kinds of

audiences might simply explain any contrasts in style. However, a specialised parallel corpus, whose contents are selected so as to balance the dates, genres, and other aspects of the plays in Shakespeare's First Folio as far as possible, helps address this in my study. Though it does not eliminate some inevitable diversity among individual authorial styles, it enables some conclusions about Shakespeare's style, in the context of the combined styles of a range of his contemporaries, to be reached.

My study does not (and could not possibly) aim for a strict definition of "Shakespearean" and "non-Shakespearean" style in EModE plays. Rather, it seeks to illuminate style features, especially those which may be hitherto undiscovered, through using statistically-based methods which point to trends in language similarities and differences between Shakespeare and other contemporaneous playwrights. I take Crystal's view that:

> firm statements about style are going to be elusive. But careful analysis can certainly identify stylistic preferences, and sometimes even a quite small observation can be intriguing. (2008:18-19)

My analyses in chapters 6, 7 and 8 demonstrate this. Space does not permit all the quantitative results to be broken down and discussed by dramatic genre, gender and/or social rank of characters who are speakers and addressees in the plays, though these factors have bearing on style. My analyses therefore provide an overall picture of Shakespeare's language style compared to that of a group of his contemporaries, which is made robust by the systematic methods applied and the attention paid to the distribution of results and evidence from the surrounding co-text and context. My study provides some initial insights into language style features which appear to be shared (or not shared) by Shakespeare and other playwrights of the period, some of which can then be followed up in more detail in future research (suggestions for which are made in 9.4). The overall comparisons enable me to perform an important

secondary task to the stylistic analysis, which is to monitor how successful my efforts (discussed in chapters 4 and 5) are in ensuring that the contents of the new *NDC* are comparable to Shakespeare's First Folio, and to assess whether the nature of the texts in the corpora biases the output from the corpus linguistic software in any way(s). This is of value to the study, to avoid drawing any mistaken conclusions about authorial styles. An overall comparison of the two corpora is also sufficient to test out and extend Baker's (2011) cutting-edge "locking" technique (chapter 7).

As indicated in the previous section, and explained further in chapter 2, my study combines elements of corpus linguistics and stylistics in its approach to EModE drama. Drama is argued by Culpeper and Kytö (2010) to be a "speech-related" text-type, and they and other scholars (e.g. Arnovick 1999; Lutzky 2009a, 2012; Lutzky and Demmen, forthcoming) use it in historical sociolinguistic research. However, limited space permits only brief links to historical sociolinguistic studies to be made in my analyses, where directly relevant to my findings. With regard to literary critical research, I share the view of other corpus linguists working with literary texts, e.g. Busse (2006:51) and Mahlberg (2007:19-20), that literary and linguistic approaches are complementary, leading to a more profound understanding of texts from two very different methodological angles. However, my engagement with the many literary critical studies of EModE plays is again necessarily restricted to those which can be closely linked with my research aims and findings. In particular, studies that make use of historical evidence, such as that of Jardine (1983), are helpful in contextualising and explaining my quantitative language data from EModE plays (e.g. in 7.4). The choice of plays to include in the parallel reference corpus (in 4.3) is also largely informed by literary critical discussions, since linguistic research into EModE plays is so scarce.

I have clarified the main aims of the study, above, and their scope, which now enables me to set out my formal research questions.

1.3 Research questions

In order to fulfil the aims of the study set out in 1.2, the following three research questions need to be addressed. Research question 1 is analytical, and research questions 2 and 3 are methodological.

1. What do similarities and differences traced through statistically significant single words, word clusters and semantic domains in Shakespeare's plays and other contemporaneous plays reveal about:
    - 1.1 The style of Shakespeare's language compared to his peers?
    - 1.2 The register of EModE drama?
2. What kinds of corpus linguistic methods best serve the investigation of similarities and differences in the language of Shakespeare's plays and other contemporaneous plays?
    - 2.1 Is the application of the locking method to words and other language features a useful addition to the application of the keyness method for stylistic analysis?
    - 2.2 Are there any issues with applying the locking method to (a) historical corpora or (b) synchronic corpora?
3. What issues arise in building a corpus of historical drama that will successfully serve as a parallel reference corpus for Shakespeare's plays?
    - 3.1 What source texts of plays that are contemporaneous with those of Shakespeare are available?

3.2 Which are the most suitable play-texts to include in a parallel reference corpus for Shakespeare's First Folio, from those that are available?

3.3 How should the texts in the corpora be prepared to better facilitate investigation using the chosen methods?

3.4 How can the inherent problems of working with texts from the 16th and 17th centuries, such as spelling variation, be addressed to render them suitable for a corpus analysis?

Having set out the rationale for the study, the approach and methodology I will use, and my aims and specific research questions, in the remainder of this introductory chapter I now give an outline of the rest of the thesis (in 1.4), followed by a list of some further terms and conventions used (in 1.5).

1.4 Outline of the study

Theory and methodological processes are closely related in this study. For convenience and clarity my discussions of theory and methodology are therefore combined, and distributed over four separate chapters. These are followed by three chapters containing my analyses.

Chapters 2 and 3 (together with later chapters containing my analyses) contribute to the answering of research question 3. In chapter 2 I discuss the concepts of "style" and "stylistics", the field of corpus stylistics, some relevant issues in corpus linguistics, the principles behind the methods used to explore the corpora, and what they will add to a stylistic analysis of EModE plays. This includes a discussion of the fairly well established method of keyness, and some criticisms, as well as the kinds of language constructs to which it has been applied. I make clear why I have chosen to include word clusters and semantic domains in preference to other linguistic constructs

(e.g. parts of speech) in my study. I also give a full account of the nature and prospective benefits afforded by the investigation of similarities as well as differences between the corpora of Shakespearean and other contemporaneous plays, and how keyness tools can be harnessed to provide the "locked" results.

In chapter 3 I discuss in more detail some issues concerning the application of the corpus linguistic methods detailed in chapter 2. I explain the concepts of word clusters and semantic domains, and what they will usefully add to my study, in view of some likely potential problems (from evidence in other research). I also set out my operational definitions for keyness and locking, by providing details of the settings and parameters used in *WordSmith* and *Wmatrix*. I explain the rationale for these, to ensure as far as possible that my methods can be replicated. In chapter 3 I also discuss the importance of distribution to the interpretation of my corpus results (in the subsequent stage of qualitative stylistic analysis), and finally I highlight the influence of reference corpora on empirical data from corpus studies. This leads me to a discussion of the compilation and preparation of the two corpora used in the study, in the subsequent two chapters. These address research question 2.

Chapter 4 concerns research questions 2.1-2.2. I begin by explaining how the *SDC* was compiled and annotated for previous research, why it is suitable for the present study, and what its limitations are. I then discuss the compilation of the new parallel corpus (the *NDC*). There are many issues and problems in the construction of a historical corpus, particularly one designed to be compared closely with an existing collection of texts (Shakespeare's First Folio). I evaluate possible sources, the choice of plays, editions used, and issues of compatibility surrounding texts from different sources. The chapter ends with a brief section giving quantitative data on the contents

of the *SDC* and the *NDC*, broken down by genre and other factors which underlie their comparability.

Following the compilation of the *NDC*, considerable treatment of the texts is required to render them suitable for successful investigation with corpus linguistic software tools. This is documented in chapter 5, which addresses research questions 2.3-2.4. I discuss the exclusion of non-dialogic text through annotation, and the need to deal with missing text and typical spelling irregularity in texts from the Early Modern period. As I explain, these factors have direct bearing on the quality of empirical results on which the qualitative analyses in subsequent chapters are based. Again, the details of chapter 5 aim to make clear the benefits and limitations of the new corpus, as well as to ensure replicability of the methods used to prepare it.

Chapters 6, 7 and 8 contain my analyses of the language style of Shakespeare in the context of the combined styles of other contemporaneous playwrights in the *NDC*. These chapters address research question 1, concerning authorial styles and register features of EModE drama. They also contribute to answering research question 2, concerning suitable methods for investigating similarities and differences in corpora (with chapters 2 and 3). Chapters 6, 7 and 8 are respectively oriented to the analysis of simple frequency, locking and keyness. My overarching approach is to present and discuss the evidence for similarity and difference between Shakespeare's language style and that of his peers. In chapter 6, I look at the most highly frequent words, word clusters and semantic domains in each corpus, when these are extracted from the corpora independently of one another. Similarities provide a logical context for differences, so in chapter 7 I then focus more closely on similarities, by examining the output of "locked" words, word clusters and semantic domains. In chapter 8, I focus on the evidence for differences in Shakespeare's language style compared to the

other contemporaneous dramatists, using keywords, key word clusters and key semantic domains.

Chapters 6, 7 and 8 all end with summary-discussions that look at the implications of the results for authorial styles, as well as any evidence for language features which are characteristic of the register of EModE drama (a distinction I discuss in 2.2). I also evaluate the output produced by each of the methods, and what value it adds to the investigations, and I comment on the impact of any problematic results which arise. Finally, in chapter 9, I provide a summary of the main findings, outcomes and conclusions of my study, an assessment of how well my research questions have been answered, and some suggestions for further research.

1.5 Definitions and conventions used in this study

1.5.1 Definitions

Unless stated otherwise, the following definitions apply, in addition to those given elsewhere in the thesis:

- Wales' (2001) definitions of stylistic terms and general linguistic concepts (for example, "co-text" refers to words surrounding other words in a text, whereas "context" indicates broader situational circumstances; Wales 2001:82, 88).
- By "discourse", I mean "a series of connected utterances, a unit of potential analysis larger than a sentence" with a "transactional" nature (Wales 2001:115). I follow Crystal (2008:208) in considering that discourse in drama encompasses that between characters on stage, as well as that between characters and audience (in the form of soliloquies).
- By "text" I mean any spoken or written "language event" (following Scott and Thompson 2001:4, who cite Hoey 1991).

- A "play-text" refers to the actual written form of a play under consideration, following Culpeper and McIntyre (2006:775); this term is a helpful reminder that the study focuses on written forms of drama, not performed forms.
- In this study, "sex" refers to a biological characteristic, whereas "gender" refers to a socially-based set of characteristics which are constructed through language and other behaviours (following e.g. Talbot 2010:7).
- Where I mention social ranks of people in the Early Modern period, I am referring to distinctions between social status levels made by, e.g., the historical sociolinguists Nevalainen and Raumolin-Brunberg (2003:28-43).
- I use Nevalainen's (2006:1) dates for the EModE period, between circa 1500 and 1700. However, the data in my study does not cover this entire period; since it is limited to the works of Shakespeare and his contemporaries, it spans only the late 16th and early 17th centuries. By present-day English ("PDE"), I mean usage from the late 20th century onwards.
- I use Baker et al.'s (2006) definitions of general corpus linguistic terminology throughout this thesis (e.g. "tags", "annotation", "reference corpus").
- I use the term "recurrent word combination" to mean "any continuous string of words occurring more than once in identical form" (Altenberg 1998:101). Other similar generic terms include "multi-word unit", "word cluster" and "n-gram", although there may be issues of compatibility among different studies. The term "word cluster" in my study is specific, and refers to the computer-generated recurrent word combination data extracted with *WordSmith*. This kind of word cluster is similar to Biber et al.'s (1999) concept of a "lexical bundle" (as argued by Scott and Tribble 2006:12, 32; see further 3.2).

- I describe the dialogue of characters in the play-texts as being arranged in "speech turns" or "speaking turns" which contain "utterances" constructed by the playwrights. The length of an utterance can be from a single word to multiple sentences (Wales 2001:401), and the concept of an utterance takes in the situational context as well as what is actually said (Levinson 1983:18-19, cited by Culpeper and Kytö 2010:8).

1.5.2 Conventions

- I show my results in capital letters in tables and occasionally in my discussions, e.g. I AM NOT, to distinguish them from general citation of the words and phrases which the results contain or comprise. General citation is shown by italics, e.g. *I am not*.
- My emphasis is indicated by boldface type.
- In my examples from the corpus texts, language which occurs as results in my data is indicated with underlining.
- I provide act and/or scene references where they are available from the play-texts in the corpora. Some of the early extant play-texts in the *NDC* are segmented into acts but not scenes, and some are not segmented at all.

CHAPTER 2. INVESTIGATING LANGUAGE STYLES IN LITERARY TEXTS USING CORPUS LINGUISTIC METHODS

2.1 Introduction

This chapter and the next explain the methods chosen to investigate language styles in plays by Shakespeare and other contemporaneous playwrights, in consideration of existing research. The content of chapters 2 and 3 underpins the answering of research question 2, concerning corpus linguistic methods for investigating my two corpora of EModE drama. In the present chapter, I discuss the methodological principles on which my findings are based. Section 2.2 concerns stylistics and style, with a particular focus on drama, and 2.3 concerns corpus linguistics. A more detailed discussion of the interface between these two areas, in the sub-field of "corpus stylistics", follows in 2.4. Here, I explain how corpus linguistic methods can aid stylistic analysis, and I note some criticisms that have been levied. In 2.5, I present evidence from other studies in support of my choice to use the methods of simple frequency and keyness in my analyses, again noting some criticisms. I explain that these methods can be applied to a variety of linguistic constructs, from which I have opted to include word clusters and semantic domains in addition to single words.

In 2.6, I go on to argue that although frequency lists and key results generate useful results that help address my research aims, a further technique needs to be included which draws attention to potential similarities in language styles. This is because keyness is oriented to differences, and similarities are also worthy of attention (Baker 2004:349), though they are often not addressed in corpus stylistic studies. I note some possible approaches to exploring language similarities between corpora, then I explain my rationale for applying and extending Baker's (2011) concept of statistical "locking" in this study, by adjusting the parameters of the keyness tools in

WordSmith and *Wmatrix*. In 2.7 I discuss some issues surrounding the interpretation of corpus data, which have implications for qualitative analysis.

My discussions mainly concern other studies of literary texts, since they are most closely relevant to my own, though the methods of simple frequency and keyness have of course been applied much more widely. Further discussions of stylistics can be found in, e.g., Jeffries and McIntyre (2010), Lambrou and Stockwell (2007), Leech (2008), Leech and Short (2007), McIntyre and Busse (2010), Semino and Culpeper (2011), Semino and Short (2004) and Wynne (2006). For more detailed discussions of corpus linguistic methods and issues in corpus linguistics, see for example Baker (2006), Hunston (2002), McEnery et al. (2006), McEnery and Hardie (2011), McEnery and Wilson (2001) and Stubbs (e.g. 1996, 2001).

2.2 Stylistics, style and the register of drama

In this section, I begin with a brief discussion of what "stylistics" encompasses, and I explain briefly how my study fits into the field. I then go on to clarify what is meant by the terms "style", "register" and "genre" in this study, and I also mention some particular characteristics of the language of drama and its historical investigation.

Broadly speaking, Leech defines stylistics as "the study of *style*; of how language use varies according to varying circumstances" (2008:54, Leech's emphasis). With regard to its application to literary texts, Leech and Short (2007:3) state that stylistics is "the study of the relation between linguistic form and literary function". As Ho (2011:5) says, "[s]tylistic analysis relies on linguistic evidence in the literary work, and thus makes use of various tools of linguistic analysis."

Semino and Culpeper (2011, citing Leech 1985) make a distinction between "general stylistics" and "literary stylistics", the former relating to non-literary texts

and the latter to literary texts. Wales (2001:373) distinguishes further between "literary stylistics", which is concerned with describing the text-type or literary genre, and "linguistic stylistics", which is concerned with a linguistic theory or model (and see also Jeffries and McIntyre 2010:2). My study concerns literary stylistics, through its overarching aim of investigating variation between authorial styles within the single text-type of EModE plays. However, I follow other scholars in using the term "stylistics" to stand for the narrower sense of literary stylistics in my discussions.

As stated in 1.2, the furthering of my main aim of investigating style in EModE drama incorporates corpus building and preparation, and the testing and extension of corpus linguistic methods, as well as a comparison of the ways in which Shakespeare and other contemporaneous playwrights use language in similar and different ways. My linguistic evidence is empirical data indicating language features that are used relatively statistically frequently, infrequently, or to a similar extent by Shakespeare and other playwrights. Analyses of their function and use in the play-texts leads me to findings about Shakespeare's authorial style, and more widely about the register of EModE drama.

My concept of "style" in this study takes in the definitions of Biber and Conrad (2009) and Leech and Short (2007). Leech and Short (2007:32) argue that the term "style" can have both "broader" and "narrower" meanings, and that it may include the following elements:

> a way in which language is *used* [...] *choices* made from the repertoire of the language [...] a *domain* of language use (e.g., what choices are made by a particular author, in a particular genre, or in a particular text) [...] *explaining the relation between* style and literary or aesthetic function. (2007:31, Leech and Short's emphasis)

Leech and Short's explanation of style accounts for the fact that authorial choices about style are always situated and context-dependent, as well as being aesthetic. This

is helpful in incorporating language choices in drama which appear to be artistic as well as functional, which I anticipated in my findings because dramatic dialogue performs a number of linguistic functions at the same time (explained further below). However, based on existing research into EModE drama such as that of Culpeper (2002, 2009), Culpeper and Kytö (2010) and Culpeper and McIntyre (2006), I also anticipated that my findings would include:

- some linguistic features which appear to be there because they have useful functions in communicating a play to an audience (or reader);
- some which construct particular varieties of drama; and
- others which seem to be personal preferences of Shakespeare and/or other contemporaneous playwrights.

To describe these more clearly, it is therefore useful to have some narrower conceptions of register, style and genre as well. Biber and Conrad distinguish between the linguistic concepts of "register" and "style" as follows:

> Register features are pervasive linguistic features that are functional; that is they are frequent because they conform to the situational context and communicative purposes of the texts in the register. Style features are similarly pervasive linguistic features, but they are not directly functional. Rather, they reflect attitudes about language, and aesthetic or artistic preferences. Thus, texts from the same register, sharing the same situational context and the same communicative purposes, can differ in their linguistic *styles*. (2009:151, Biber and Conrad's emphasis)

This distinction is not clear-cut, however, especially in dramatic dialogue constructed 400 years ago. Even with a relatively large amount of data, it is difficult to say whether some kinds of language trends are more accurately described as style choices made by more than one dramatist (i.e. based on social or personal preferences for ways of expressing meanings, or constructing pragmatic or rhetorical strategies), or register features (choices made to enable a late 16th/early 17th century audience to engage

with the play). They can arguably be both, given the multi-functional nature of dramatic dialogue, mentioned above.

The language in plays by the dramatists whose work is included in my study will have been constructed to achieve not only communication between characters at an on-stage level, but also the conveyance of essential information to the audience about what is going on in the world of the play. It is helpful to think of this dual function of dramatic dialogue in terms of Short's (1996:163-172) concept of "discourse architecture", in which on-stage communication between characters occurs at a lower "discourse level", and communication from playwright to audience (or reader) is transmitted at a higher discourse level[5]. This layered structure is shown in Figure 1.

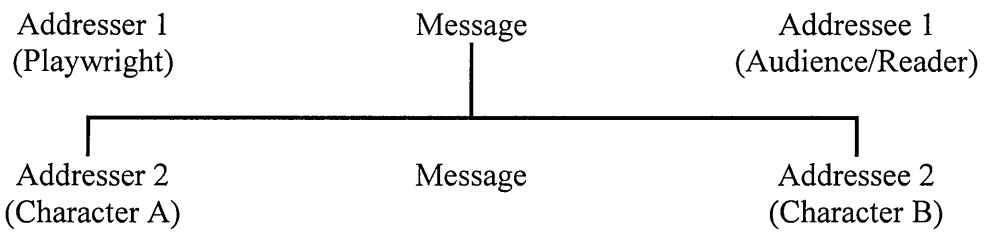


Figure 1. Discourse levels in drama (from Short 1996:169)

Short (1996:169-170) explains that through the higher discourse level, the playwright can provide the audience with a privileged understanding of what is happening on-stage, since the audience hears and sees things which some of the on-stage characters do not. It is at this higher discourse level that important dramatic functions such as characterisation, suspense, irony and comedy make their impact. This makes it difficult to separate language that constitutes an aesthetic choice (qualifying as a style feature in language) from that which is an entirely functional choice associated with the communication of the play (a register feature). Petersen (2010:xviii) considers the

[5] There may also be a third level, if the play has a narrator, although none of those in my corpora are narrated.

style features evidenced by the grammar patterns in her EModE drama data to be on a cline, describing them as relatively more or less "formulaic" or "general" at one pole, and relatively "authorial" at the other. This is also a helpful approach to discussing the register and style features which emerge from results in my analyses, rather than as two discrete categories.[6]

I use literary concepts of genre that were associated with drama at the time the plays were written, i.e. those of comedy, tragedy and history (though as discussed further in chapter 4, they are neither ideal nor the subject of consensus). A linguistic concept of genre, such as that of Biber and Conrad, would be harder to apply, since it encompasses a "culturally expected way of constructing texts belonging to the variety" (2009:16). The historical gap between the Early Modern period and the present day inevitably creates a gap in cultural understanding (through changes in, for example, philosophical and theological thought, and ideas about language, as discussed in e.g. Hope 2010). The plays in this study were written at a time when a shift from religious authority towards human and scientific authority was just beginning (see e.g. Hillman 1997; Hope 2010:35-37), and the ways people thought about themselves and interpreted the world around them are difficult to appreciate fully from a 21st century perspective. Hope (2010:31) argues that the art of self-expression in Renaissance drama is not based on the delivery of characters' original personal insights, such as we might assume today, but rather on the incorporation of parts of existing texts that were common knowledge, and part of "a shared stock of ideas on which everyone drew".

Historical sociolinguistic research into EModE (for example, Nevalainen 2006 and Nevalainen and Raumolin-Brunberg 2003) is of course helpful in providing

[6] Petersen's (2010) results are not comparable to mine, however, since I am not examining grammatical features in my data, and my aims do not include the attribution of authorship, as hers do.

contextual information about drama of the period. However, as Mazzon points out with regard to medieval drama:

> No matter how much we know about the social context and the writing conventions of distant epochs, there are always serious doubts as to the perlocutionary effects, and the degree of felicity, of any communicative act when it was originally performed. (2009:2)

This is perhaps true to a slightly lesser extent of EModE drama, since the language is in many ways similar to PDE (Crystal 2008:230), and a social rank structure with a middle section of merchants and traders, more familiar in comparison to that of today, was by then emerging (see Nevalainen and Brunberg 2003:28-43). Audiences of plays performed for public theatre such as those in this study (see 4.3.2.3) were drawn from all social ranks, and their cultural expectations may have varied.

The above points not only explain the difficulties with using anything other than a literary concept of genre in this linguistic study, but also highlight some general considerations to bear in mind when analysing historical drama. Furthermore, it is also worth remembering that although Shakespeare and other contemporaneous playwrights would have been familiar with the concepts of grammar and rhetoric (see e.g. Hope 2010:30-37), the dialogue in their works would have been constructed without any explicit idea of linguistic concepts such as word clusters and keywords, used to describe and analyse it in studies such as mine. These belong to the present-day field of corpus linguistics, which I now discuss in more detail.

2.3 Issues in corpus linguistics which are relevant to this study

There are some potentially ambiguous and/or contentious terms in corpus linguistics, some of which I now discuss briefly to clarify my own position, beginning with what qualifies as a "corpus". Corpus linguists such as Baker (2006) and McEnery et al. (2006) describe a "corpus" as a range of texts sampled from multiple choices

available, according to certain criteria and principles (and see also McEnery and Wilson 2001). Mahlberg and Smith (2010:449) state that "[t]he term 'corpus' refers to a collection of computer readable texts. Corpora are normally large, that is, containing many millions of words." The 100 million-word *British National Corpus* (hereafter the "*BNC*"; see Aston and Burnard 1998 and Burnard 2000)[7] is an example of a large, prototypical "general" corpus, which is used to investigate language patterns across a broad variety of text-types. The application of corpus linguistic methods to specific text-types has seen a trend in recent years towards building "specialised" corpora, which Baker (2006:26) and McEnery et al. (2006:13-19) define as those that are designed to answer a specific set of research questions. An example is the one-million-word *A Corpus of English Dialogues, 1560-1760* (hereafter the "*CED*"; see Culpeper and Kytö 2010; Kytö and Walker 2006): a diachronic corpus containing a range of text-types from multiple registers of historical speech-related areas including drama (Culpeper and Kytö 2010:25).

Corpora of literary texts tend to be specialised corpora and, as Mahlberg and Smith (2010:449) point out, they are typically much smaller than general corpora. They may feature just one electronic text (e.g. a play or a novel), chosen for investigation on the basis of its perceived worth, such as its cultural or literary interest. Existing specialised corpora of literary texts vary considerably in size and source(s). Some comprise a text or texts by a single author, for example:

- Mahlberg's (2007:6) corpus of Charles Dickens' novels (c. 4.5 million words);
- Fischer-Starcke's (2009:497) corpus of Jane Austen's novels (c. 735,000 words);

[7] See http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro (accessed 19.05.12).

- B. Walker's (2010:365, 2012) corpus of Julian Barnes' novel *Talking It Over* (c. 73,000 words);
- Inaki and Okita's (2006:285) corpus of Lewis Carroll's two novels *Alice in Wonderland* and *Alice Through the Looking Glass* (c. 56,000 words); and
- Culpeper's (2002:15) corpus of the dialogue of six characters in Shakespeare's *Romeo and Juliet* (c. 20,000 words).

Others contain samples of texts from multiple authors, e.g.:

- Semino and Short's (2004:19) corpus of 120 samples of approximately 2000 words from late 20th century British English fiction, newspaper reportage and biography/autobiography, totalling 258,348 words; and
- Mahlberg's (2007:6) corpus of 19th century British fiction (comprising about 4.5 million words from 29 separate texts by 18 different authors), which she uses as a reference corpus for her abovementioned Dickens corpus.

On the basis of the above, I describe the two collections of EModE plays in my study as "specialised corpora". As detailed fully in chapter 4, one is a single-author collection (the *SDC*, comprising 36 plays by Shakespeare), the other is a multiple-author collection (the *NDC*, comprising 43 plays by a total of 23 other contemporaneous playwrights), and both are about 800,000 words in size. The *NDC* is arguably a more prototypical corpus than the *SDC*, since its contents are sampled from a range of other available works, whereas the *SDC* contains all that is available of its kind (plays comprising Shakespeare's First Folio). However, the "samples" in the *NDC* are whole plays, which vary somewhat in size, rather than excerpts of similar length, such as are in Semino and Short's (2004:19) corpus. My approach of using a single-author corpus and a reference corpus representative of a wider range of texts from the same literary genre and period is similar to that of Mahlberg (2007:6), mentioned

above, and also to Leech (2008:167), who uses a reference corpus comprising samples of other approximately contemporaneous fiction by multiple authors for his keyness analysis of Virginia Woolf's short story *The Mark on the Wall.*

Not all scholars would agree with my above rationale. Louw (2008:254) criticises the use of the term "corpora" for stylistic studies of small bodies of text, on the basis that these may not be sufficiently representative of the language areas from which they are sampled. Louw prefers the term "digital" stylistics, and other alternative terms have been suggested for studies of smaller and/or less prototypical collections of text. These include "electronic text analysis" (Adolphs 2006:1-2); "computer-aided", "computer-assisted" or "corpus-assisted" (Balossi 2009:7, 78, 79); and "small-corpus-based" (Inaki and Okita 2006). Whilst accepting some difficulties and limitations of using the term "corpus" in a broad way, I would argue that substituting an alternative does not necessarily make the contents of an electronic collection of texts, or how it is treated, prepared and exploited, particularly transparent. For example, it is easy to think of examples of a relatively large corpus, such as the *BNC*, and a relatively small one, such as that of a single novel, but hard to say where the cut-off point between "large" and "small" would lie, or what would be the exact sampling criteria for a "corpus" to qualify as such. It therefore seems better to use the umbrella term "corpus" to refer to collections of electronic text(s) in a wide sense, but also to be explicit about the purpose of the corpus and how it is constructed and compiled. In the case of a collection of text samples, it is vital to make clear the selection principles that are followed, hence the devotion of a whole chapter in this study to the construction and contents of my two corpora (chapter 4).

I also use the term "corpus-based" broadly, to mean research based on empirical results derived from electronic texts by applying corpus methods. I take

Wynne's (2010:426) view of concentrating on what can be learned about language through the use of corpora and corpus tools, rather than on debating distinctions between methodological and theoretical aspects of corpus linguistics[8]. Again, it is important to clarify the principles and practice of methodology, however. In the field of corpus linguistics, there is a distinction between "corpus-based" and "corpus-driven" approaches, made by Tognini-Bonelli (2001). In this view, corpus-based research involves results selected according to categories which are pre-chosen by the researcher, whereas corpus-driven research involves results which are categorised according to what the data yields (see also Baker 2006:16). In recent discussions of the two terms, Biber (2009) points out that research may in fact be both corpus-based and corpus-driven, and Rayson (2008:519) conceives of the two approaches being blended into a "data-driven" approach, in which statistically-based results emerge from the texts and guide the researcher to the most potentially interesting language features for analysis. My study can be considered data-driven, because all my analyses proceed from quantitative results arising from the statistical processes applied.

A related issue is the orientation of the study: from bottom-up or top-down. A bottom-up approach takes empirical language data from the text as a starting point, whereas a top-down approach begins with some pre-determined categories or assumptions, and is more intuition-driven (see further Archer 2009:12, FN 30; Jeffries and McIntyre 2010:12-14). The data-driven aspect of my study means that my overall approach is bottom-up. However, as discussed further in 2.7, two categorisation frameworks are imposed on my data (functional categories for word clusters; see further 3.3.2.5, and semantic domain groupings; see 3.3.1). These add a top-down element.

[8] See Wynne (2010) and other papers in the *International Journal of Corpus Linguistics* 15(3) for a recent airing of this debate.

The matters discussed in this section relate to the general discipline of corpus linguistics, not just to its application in stylistic studies. Next, I discuss issues surrounding the interface between corpus linguistic and stylistic investigations.

2.4 Corpus stylistics: the area of this study

In this section I begin with some definitions and principles of "corpus stylistics", in 2.4.1, followed by a discussion of its advantages and criticisms, in 2.4.2. Lastly, in 2.4.3, I clarify the distinction between corpus stylistics and the related area of computational stylistics, in which research into Shakespeare's plays and other EModE drama is also being carried out (as indicated in 1.1).

2.4.1 Defining corpus stylistics

Corpus stylistic research constitutes part of the relatively new field of "digital humanities", mentioned briefly in 1.1, which emerged with late 20th and early 21st century advances in computer technology and electronic file storage (see further e.g. Galey and Siemens 2008). Corpus stylistics essentially involves the use of corpus linguistic software tools to extract quantitative data from electronic collections of literary text(s) using statistical methods, the results of which can help the researcher to pinpoint markers of language style, and language features which create particular stylistic effects. Language of stylistic interest is argued by Leech (2008:55) as occurring in the form of "deviations" from a relative norm. Corpus linguistic software tools offer the potential for establishing norms and identifying language which departs from them on a statistical basis, by counting and displaying "repeated and typical uses" (Mahlberg 2007:4) which the researcher can then investigate further. In discussing language which "stands out", it is helpful to make a distinction between

that which is statistical and that which is psychological. Leech and Short (2007:39-41) distinguish between "deviance" (a statistical, quantitative concept) and "prominence" (a psychological, qualitative concept). I prefer to use the term "quantitative significance" to describe numerical frequency and/or statistical significance in this study, because the term "deviance" seems to imply difference, which makes it an awkward descriptive term for results indicating similarities in my data. I use Leech and Short's term "prominence" to mean language that stands out psychologically (to me, with the caveat that this may be subjective).

Leech and Short (2007:40) "presume a fairly direct relation between prominence (psychological saliency) and deviance (a function of textual frequency)", although they emphasise that that not all language features which are statistically unusual are also psychologically noticeable (i.e. foregrounded, a concept mentioned in 1.1). This is because readers (and, by extension, audiences of plays) do not all notice the same features, and some features may not be frequent enough to be noticeable (regardless of qualifying as statistically significant in a set of data). In other words, quantitatively significant language is often also psychologically prominent, but not always. Corpus linguistic methods can help find language style features which are not psychologically prominent, by identifying items which are quantitatively significant, as I discuss further in the next section.

2.4.2 The advantages of corpus stylistics and some criticisms

There are three main advantages to using corpus linguistic methods for stylistic analysis:

(i) Language features and patterns which would be difficult or impossible to spot through manual analysis, or by using intuition alone, can be highlighted using

statistical processes (argued by, e.g., Fischer-Starcke 2009:494; Ho 2011:6-7; and McIntyre 2010:169).

(ii) Automating the identification of potential deviations in language is more systematic. It reduces the risk of subjectivity inherent in relying solely upon intuitive choices about what kinds of language to analyse in literary texts (as argued by, e.g., Ho 2011:6-7; Mahlberg 2009:48; O'Halloran 2007:228 and Stubbs 2005:22).

(iii) Corpus linguistic software tools facilitate the rapid counting and statistical comparison of words in much larger quantities of text than could feasibly be analysed by human effort alone (stated by, e.g., Leech and Short 2007:286 and McEnery et al. 2006:5-6).

Despite the above advantages of using corpus methods to investigate literary texts, there have been a number of criticisms (see Archer 2007; Ho 2011:9-11; Jeffries and McIntyre 2010:22, 181-182; Mahlberg 2009:48; Wynne 2006:225-226). These can be summarised as follows:

(i) Corpus investigations are the result of a subjective and circular process, since the output is pre-determined through finding language features which have already been chosen as interesting. The most well-known critic taking this stance is probably Fish (e.g. 1980, 1996, 2012).

(ii) Corpus methods risk reducing literature to a decontextualised, numerical list of language features (pointed out, for example, by van Peer 1989:302-305).

Responses to Fish have been made by Craig (2004:277-278), Stubbs (2005:6) and Witmore (2012), amongst others. More generally, stylisticians who use corpus linguistic methods argue that a systematic approach and careful interpretation of results can greatly reduce the risks of subjectivity and circularity (see for example Ho

2011:9-11 and Mahlberg 2009:48-49). In my study, the data-driven approach itself, discussed in the previous section, helps avoid circularity. This is because the computer identifies a limited set of potentially interesting results, on a statistical basis, which vastly constrains the element of personal choice about what to analyse qualitatively. The decontextualisation of language features, when they are presented as output in computer-generated lists, can in fact be helpful at the start of a qualitative analysis. It lessens the potential for introspective choices about what to follow up, by presenting all quantitatively significant items (including those which may be from well-known speeches or characters), without their surrounding "contextual baggage"[9].

However, as McEnery et al. (2006:7, citing Leech 1991:14) point out, human intuition is still required to analyse quantitative results in a useful way. It is necessary to distil the computer-generated results which are genuinely linguistically interesting from those which are not, and to demonstrate their value through careful qualitative analysis. This is argued by corpus stylisticians including Culpeper (2009:39-40), Ho (2011), Mahlberg (2009:62) and Semino and Short (2004). As Ho (2011:10-11) emphasises, "[q]uantification and statistics should always be utilized as a means rather than an end, to verify or refute our intuition-based analysis."

In the end, no methodological approach is perfect. I concur with the view of other corpus stylisticians who uphold the benefits that corpus linguistic methods brings to the investigation of literary texts, whilst acknowledging that it is an approach which complements (rather than replaces or supersedes) others, such as those in the literary critical discipline (see e.g. Hope 2010:386-387; Lambrou and Stockwell 2007:3; Mahlberg 2009:50; Semino and Short 2004:7-9). It is also worth noting that past criticisms have beneficially resulted in new attempts to strengthen corpus-based

[9] I am grateful to Karen Donnelly for suggesting this useful descriptive term.

approaches to literary texts. As Jeffries and McIntyre argue, stylisticians have expanded theoretical bases and frameworks for analysis through which corpus results are analysed, now taking in cognitive approaches and "functional and discourse-analytical approaches to language" (2010:22). This helps ensure that the data are not simply extracted and then abstracted from the texts, but are instead interpreted with a close eye on the context and what may be going on between text and reader/audience.

On the basis of the above discussions, using corpus linguistic methods in my study enables a much more objective, rigorous and systematic comparison of the language style of Shakespeare and a range of his peers than would be within my manual resources. My quantitative data, and the outcomes and conclusions drawn from a close and detailed qualitative analysis of them, are based on **all** the dialogue spoken by characters in the plays investigated (i.e. about 1,600,000 words of EModE dramatic dialogue; see 4.4), rather than on selected extracts and/or the speeches of just a few characters. That is an inherent limitation faced by researchers using exclusively qualitative methods. My corpus stylistic study is concerned with the relationship between linguistic forms in EModE plays and the meanings these forms have in the construction of language styles of characters. Authorial styles are reflected in repeated choices of language forms with particular meanings and functions in:

- the construction of characters and plots;
- the creation of dramatic atmospheres (e.g. humour or suspense); and
- the packaging of dialogue in a way that also communicates a coherent story which the audience can understand and find engaging.

The route to finding the most potentially interesting stylistic features is the quantitatively significant linguistic items identified by the corpus tools, which I discuss further in the next section. First, however, I further clarify my approach by

differentiating between "corpus stylistics" and "computational stylistics", as both involve the application of statistical methods to investigate literary texts, and there are some substantial existing studies of EModE plays in the computational stylistics area.

2.4.3 Corpus stylistics and computational stylistics

The statistical processes and methods of corpus stylistics and computational stylistics tend to be different, reflecting different underlying aims. Mahlberg and Smith (2010:450) comment that "[i]n contrast to some of the more computational approaches to stylistics, such as stylometry, one may want to see methods in corpus stylistics mainly as complementing interpretation and detailed manual analysis". Computational stylistic studies of EModE drama centre mainly on authorship attribution (e.g. Craig 2010; Craig and Kinney 2009a:8), with aims of quantifying language features to reach firmer conclusions about longstanding questions of collaboration and authorship[10]. Craig and Kinney (2009b:18) apply t-tests and Zeta words to compare authorial styles of Shakespeare and other contemporaneous playwrights (building on the work of Burrows e.g. 1987, 1992, 2007; see also Craig 2004 and Hoover 2010:260-265). Petersen (2010:169-192) uses other stylometric tests such as principle component analysis and discriminant analysis in attributing authorship of EModE plays through function words and grammar patterns. Argamon et al. (2007:811-813) use "machine learning" processes to identify and count language features typically associated with male and female characters in Shakespeare's plays[11].

The main focus of the above studies is on quantitative statistical processes, and how the resulting language patterns profile particular authorial styles, rather than on investigating their effects in the plays. An exception is Burrows (1987:3, 34-45), who

[10] Grieve (2007) provides a fairly recent overview of authorial attribution studies.
[11] For an overview of machine learning processes, see Jockers and Witten (2010).

shows that different character identities in Jane Austen's novels are shaped by the relative frequency of use of different pronouns and other function words. In contrast to computational stylisticians, corpus stylisticians tend to use quantitative data as a starting point. They make use of software tools which operate independently of the texts being investigated (such as *WordSmith* and *Wmatrix*, used in this study). These tools require less computational effort by user, though an understanding of the implications of the statistical tests and processes is imperative (see further chapter 3). The resulting quantitative data is then subjected to close qualitative analysis, carried out manually, but using further electronically-derived information about the co-text and context of the results (such as collocations and concordances) to aid interpretation.

Since the methods, statistical processes and aims are so different, it is not possible to compare my quantitative results directly with those of computational stylisticians. However, it is useful to mention occasional instances where my findings about Shakespeare's authorial style appear to coincide with theirs, especially as there is so little existing corpus-based research which compares Shakespeare's plays with other plays of the same era. Some of the distinctive words identified in Shakespeare's plays by Craig and Kinney (2009b:38) also occur in my key results, in 8.2. More generally, the evidence from computational stylistic research which indicates that Shakespeare's vocabulary is very similar to that of other contemporaneous playwrights, noted in 1.1, helps paint an existing picture of language styles in EModE plays against which my own research can be set. For example, Rosso et al., whose findings are based on computational methods of Information Theory, state that:

> William Shakespeare's plays generally use vocabulary items at a rate which is very close to the norm for the drama of his time [...] his work is unusual if anything for its constant closeness to the average use of words at the time. (2009:925)

Furthermore, Craig (2012:5) puts forward evidence indicating that Shakespeare's language style tends towards being more conservative than that of his contemporaries, particularly those who were younger. He argues that Shakespeare uses "more *hath* and fewer *has*, more *thou* forms and marginally fewer *you* forms, and more of the conjunction *that*, than his contemporaries". This chimes with the findings in Ingram and Ingram's (2012) empirical research (although it was conducted by manual analysis rather than with computers), which indicates that Shakespeare uses more older forms of syntactic constructions than his contemporaries, particularly in his later plays.

I have explained the corpus stylistic approach, above, and I now discuss the specific corpus methods which I apply to locate quantitatively significant language features of potential stylistic importance. In the next section I explain my choice to use frequency counts, keyness and locking with single words, word clusters and semantic domains in the pursuit of a greater understanding of Shakespeare's style. This addresses research question 2, concerning appropriate methods for investigating language style. It also underpins the actual findings concerning language styles in EModE plays, which are the subject of research question 1.

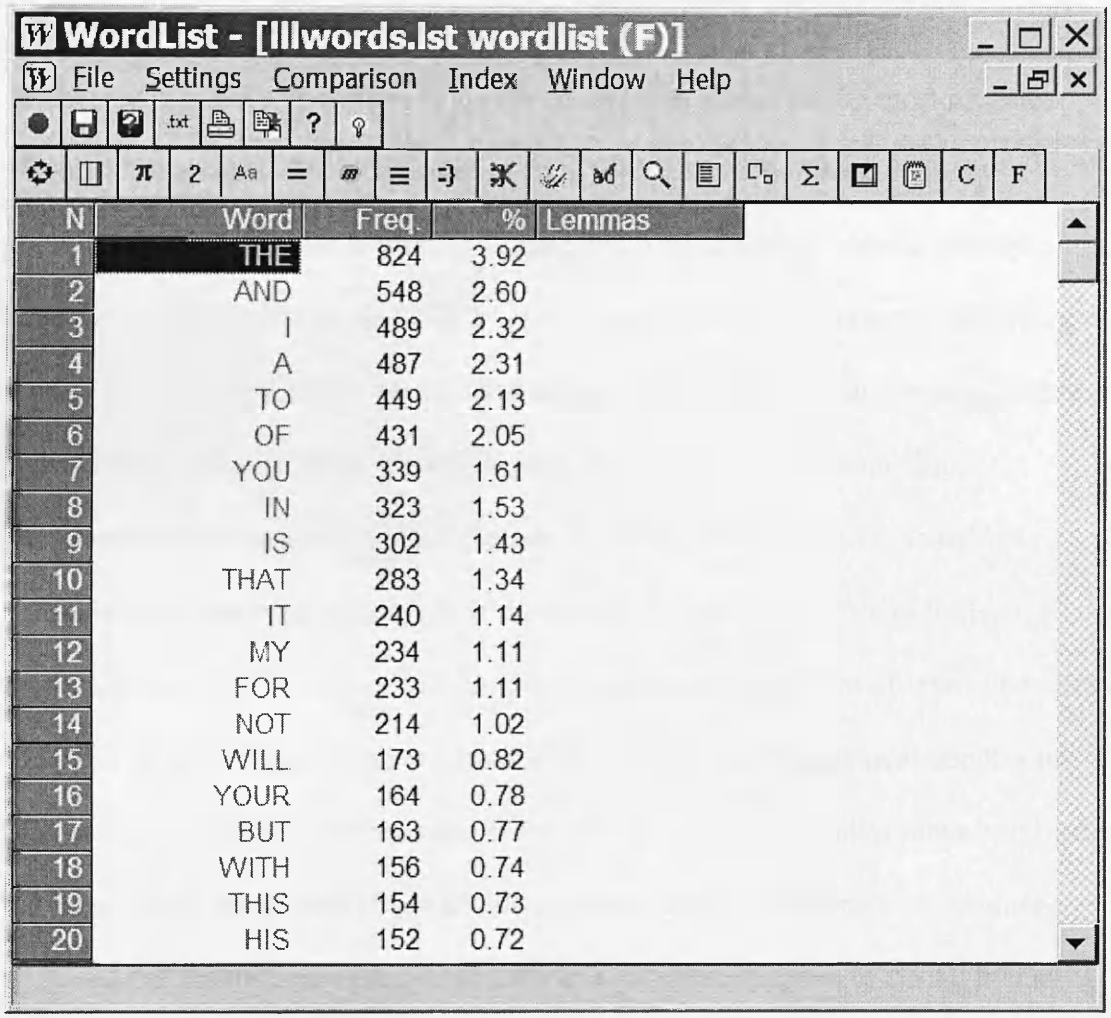
2.5 A brief review of some existing corpus stylistic methods

2.5.1 Frequency

Stubbs (2005:11) argues that "frequency lists are one essential starting point for a systematic textual analysis". As Archer explains, this is because:

> the frequency with which particular words are used in a text can tell us something meaningful about that text and also about its author(s) – especially when we compare word choice/usage against the word choice/usage of other texts (and their authors) [...] we learn something about texts by focussing on the frequency with which authors use words precisely because their choice of words is seldom random. (2009:1)

Simple frequency, therefore, is a useful place to begin comparing the language styles of Shakespeare and other contemporaneous playwrights, and the choices and preferences they made in constructing dramatic dialogue. Lists of words ranked according to frequency can be obtained from a text or set of texts using the Wordlist function in *WordSmith*[12], as shown in the screenshot in Figure 2 (which lists the most frequently-occurring words in Shakespeare's comedy *Love's Labour's Lost*).



N	Word	Freq.	%	Lemmas
1	THE	824	3.92	
2	AND	548	2.60	
3	I	489	2.32	
4	A	487	2.31	
5	TO	449	2.13	
6	OF	431	2.05	
7	YOU	339	1.61	
8	IN	323	1.53	
9	IS	302	1.43	
10	THAT	283	1.34	
11	IT	240	1.14	
12	MY	234	1.11	
13	FOR	233	1.11	
14	NOT	214	1.02	
15	WILL	173	0.82	
16	YOUR	164	0.78	
17	BUT	163	0.77	
18	WITH	156	0.74	
19	THIS	154	0.73	
20	HIS	152	0.72	

Figure 2. Screenshot showing an example of a *WordSmith* wordlist

Wordlists can also be extended to recurrent word combinations, as discussed further in 3.2. The tools in *Wmatrix* work similarly, and produce frequency lists of words, parts of speech and semantic domains (categories containing words which are related

[12] Alphabetically-ordered word lists are also produced by *WordSmith*.

semantically, discussed further in 3.3). *WordSmith* and *Wmatrix* count the number of recurrences of linguistic items on the basis of orthographic matching. This is a thorny issue when working with historical texts from a time before spelling was standardised, as I discuss much further in 5.4. For more detailed discussions of frequency lists, see Scott and Tribble (2006:11-32, 55-88), and the explorations of various frequency-based aspects of corpus linguistic methods in Archer's (2009) edited volume.

Comparisons of frequency lists provide some useful empirical evidence of how the language in texts or corpora is similar or different, based on the most prevalent linguistic constructs which characterise them. This is shown, for example, by Mahlberg (2007) in prose fiction, by Culpeper (2011) in EModE drama, and by Culpeper and Kytö (2010:116-117) in a wider range of EModE registers. Mahlberg (2007) finds that different types of word clusters characterise the dialogue of different individuals in the novels of Charles Dickens, helping to create memorable personalities and character types. Culpeper and Kytö (2010:103-141) show that different speech-related registers of EModE (e.g. drama and courtroom trials) are characterised by different kinds of formulaic language (in the form of lexical bundles). Culpeper (2011) compares the top 10 most frequently-occurring lexical bundles in Shakespearean drama with those in a much larger corpus of EModE drama (see further 4.3.1) and finds some similarities and differences between Shakespeare's language style and that in other contemporaneous drama (discussed further in 2.5.4). Including high-frequency results in my investigations builds on Culpeper's (2011) findings and takes them much further.

The above discussions do not intend to imply that low-frequency results are of no interest. However, in this study they would be less useful in obtaining a picture of Shakespeare's style in the context of other playwrights. This requires evidence of the

"repeated and typical" usage mentioned by Mahlberg (2007:4; see 2.4.1). Having shown that evidence from other research indicates that an initial analysis of high-frequency language features is worthwhile, I now explain what the keyness method will add.

2.5.2 Keywords and keyness

As indicated in 1.1, in corpus linguistics a "key" result is a linguistic item which occurs with greater or lower statistically significant frequency than would be expected in one text (or group of texts), when compared to another. Positive key items are language features occurring with comparatively high frequency, and negative key items are those occurring with comparatively low frequency (Scott e.g. 2000; see also Baker 2004; Bondi 2010:1-14 and Culpeper 2002, 2009).

In *WordSmith*, the "KeyWords" function compiles a list of keywords by comparing the contents of frequency lists, discussed in the previous section, on the basis of either the log-likelihood or chi-square statistical test (see further 3.4.1). The key results are displayed in a list ordered by statistical significance (i.e. their keyness values). Parameters such as minimum observed frequency and p-value, which determine what will be counted as key, can be adjusted by the user (discussed further in 3.4.2-3.4.3). *Wmatrix* works in a similar way (see Rayson 2008, 2009), and computes keyness on the basis of the log-likelihood test only.

The comparative advantage of a list of **key** items, compared to a frequency list, is that it enables the researcher to see what is distinctive in a text **in relation to other texts** (Baker 2006:125; Stubbs 2005:11). Keyness, therefore, offers an additional layer of statistical relativity over and above that in a comparison of the most frequent language features in two corpora. Keywords and other key results can be

indicative of quantitatively significant language features which are stylistically important[13]. Both positive and negative key results provide potential routes to finding out how language styles are characterised, through relative over-use or under-use of certain kinds of linguistic items. Consequently, the keyness method has become an increasingly popular choice in corpus stylistic studies. For example, Ho (2011) uses key semantic domains to assist her comparison of the language styles of the 1966 and 1977 versions of John Fowles' novel *The Magus*, and B. Walker (2010, 2012) uses keywords and key semantic domains to compare and contrast the language styles of the three protagonists in Julian Barnes' novel *Talking it Over*. Mahlberg (2007) uses key word clusters to investigate character construction in Dickens' novels, and McIntyre (2010) investigates the characterisation of the robbers in Quentin Tarantino's film *Reservoir Dogs*, using keywords, key semantic domains and n-grams.

The potential for keyness to aid the stylistic investigation of Shakespeare's plays, in particular, is illustrated by Culpeper's (2002, 2009) study of keywords in Shakespeare's *Romeo and Juliet*. He shows how six characters are constructed differently through what each tends to say relatively frequently in his or her dialogue, when compared statistically to the dialogue of all the other five characters. Culpeper (2002:17) notes that some distinctions can be made from the frequency lists alone (for example that the Nurse is the only character for whom the interjection *O* is one of the most frequently-used words), but that other high-frequency words are used by all or most of the characters. The pronoun *I*, for instance, is among the top 10 most frequently-used words for five out of the six characters, but that is likely to be a general feature of dramatic dialogue rather than a style marker for those particular characters. The keywords provide more evidence showing how characters' language

[13] Culpeper (2009:32) links keywords in corpus linguistics with Enkvist's (e.g. 1964:29, 34-35) much earlier concept of "style markers", distinguishing them from Williams' (1976) concept of culturally-derived keywords. See also Stubbs (2010) for a recent discussion of different "keyword" concepts.

styles contrast: Romeo's role as a lover is reflected in his relatively greater use of words surrounding the concept of love, and Juliet's state of anxiety is reflected in her comparative over-use of *if* and *yet* (Culpeper 2002:20). The Nurse has a relatively emotional language style, which Culpeper (2002:21) shows is achieved through "surge features"[14], identified through keywords such as *god*, *warrant*, *faith*, *marry* and *ah*.

Culpeper's (2002) study, and its (2009) extension to key parts of speech and key semantic domains, demonstrates that the keyness method is useful for investigating variation in language styles of characters within a single Shakespearean play. Similarly, Archer and Bousfield (2010) carry out a pragma-stylistic study of the relationships between King Lear, his three daughters and his adviser the Earl of Kent through keywords and key semantic domains in their dialogue (in the play *King Lear*). Other studies make use of keyness to investigate variation among Shakespeare's plays, for example by date, genre or gender. Murphy (2007) compares the style of language in soliloquies in early and late plays and in different genres, using keywords, key parts of speech and key semantic domains. Scott and Tribble (2006:59-70) use keywords to profile the language in *Romeo and Juliet* compared to Shakespeare's plays overall, and they find, amongst other things, that female characters in this play make relatively greater use of the exclamations *Oh* and *Ah*. Their research adds a further dimension to Culpeper's (2002, 2009) findings regarding specific characters in the play.

Despite the above evidence that keyness is a useful way of investigating language styles within and among Shakespeare's plays, existing studies have not thus far extended keyness to comparisons with plays by other dramatists from the same historical period. This approach has been used to gain insight into prose fiction by some well-known authors, however. Leech's (2008) keyness study of Woolf's *The*

[14] Taavitsainen (1999:220) argues that "[s]urge is a salient quality of personal affect in early modern fiction", and she describes surge features as language in which "[t]he speaker's or narrator's mental afflictions or temporary states of mind find linguistic outlets".

Mark on the Wall and Mahlberg's (2007) investigation of Dickens' novels both include comparisons with corpora of other contemporaneous fiction, as mentioned in 2.3 above, as do the keyness studies of Jane Austen's novel *Pride and Prejudice* by Fischer-Starcke (2009) and Mahlberg and Smith (2010). Mahlberg and McIntyre (2011) use the fiction section of the *BNC* as a comparator for Ian Fleming's novel *Casino Royale*, from which key results enable them to add empirically-based evidence to existing (qualitative) claims about Fleming's authorial style, and the characterisation of his most famous hero James Bond. These studies show that the keyness method can aid the investigation of what can be considered external variation in language style, i.e. distinctions in one author's style when compared to a range of others, as well as to internal variation between characters or works by a single author. As stated in 1.1, my study therefore helps address the gap that currently exists for a keyness study of Shakespeare's plays in comparison to other contemporaneous plays.

I have confined my above discussions mainly to keyness studies of literary texts, though the keyness method is also used to investigate many other text-types (see e.g. Baker 2009; Jeffries and Walker 2012; McEnery 2009). Despite the continuing popularity of keyness, and the weight of research indicating that it is a useful way of accessing language features of potential interest in literary other texts, like most methods it inevitably also attracts some criticism. I discuss this briefly next.

2.5.3 Criticism of the keyness method

Mike Scott (2010), the developer of *WordSmith*, has written extensively about the principles and practice of keyness in corpus linguistics, and he acknowledges that there are contentious issues surrounding the understanding and interpretation of what keyness really implies. These arise, at least in substantial part, because keyness in

corpus linguistics is a quality that only belongs to a linguistic construct when it is viewed in a certain way: on the basis of a particular statistical process, using a specific set of computer algorithms, with selected text(s) as a comparator. Keyness is, therefore, context-dependent, as Scott emphasises (2010:56). This is why it is important that statistically-derived results are used as a guide to what to analyse qualitatively, not considered to be the end product with a specific merit in their own right (as argued further in 2.7 with regard to the interpretation of quantitative data). As Scott (2010:56) says, keywords "are pointers, that is all". That is true regardless of the statistical measure or metric used to determine keyness.

Gabrielatos and Marchi (2011) and Wilson (2011) argue that the basis of measuring keyness according to statistical significance of frequencies in the corpora (i.e. using the log-likelihood or chi-square test, as noted in 2.5.2; see further 3.4.1) is flawed. They indicate that there is a lack of understanding of what the p values (the measures of probability that results do not occur by chance; see 3.4.3) actually represent about results generated from corpora, and they propose alternative bases for keyness. Gabrielatos and Marchi (2011) use the size of the effect of the frequencies of a linguistic item in the corpora that are being compared as a metric for keyness, and they demonstrate a test for this using *Microsoft Excel* spreadsheets (the "%DIFF" test). Wilson (2011), on the other hand, uses "Bayes Factors" (a concept from the field of Bayesian statistics; see e.g. D. Berry 1996). Wilson claims that Bayes Factors:

> allow the corpus linguist to quantify explicitly the degree of evidence against the null hypothesis, which is seemingly what researchers are seeking to do when they misinterpret frequentist p-values as measures of evidence for or against chance differences in proportional frequencies (2011:5)

Wilson (2011:4) applies Bayes Factors to key results generated in *Wmatrix*, using the software *R* (Ihaka and Gentleman 1996). I do not test either method in my study.

Regardless of the abovementioned criticisms of keyness based on statistical significance testing and p values, this method has been applied in many existing studies. As indicated by the findings from them which I discussed in 2.5.2 and 2.5.4, it has led to the discovery of much new and useful information about language in literary texts. I would consequently contend that, notwithstanding the arguments in favour of various measures of the actual quality of keyness, it is the interpretation of the linguistic items themselves which are highlighted as key that is the over-riding factor in the quality of any given analysis. Arguments may be made in favour of different statistical processes, but ultimately they still represent particular points of view, albeit based on sets of systematic mathematical calculations. Individual researchers need to decide which ones will best suit their aims, from the resources available to them. As Scott points out, in practice there is unlikely to be consensus over what should be considered "key", and "[k]eyness is therefore somewhat subjective, anyway" (2010:46).

This is not to say that attention need not be paid to the principles of good statistical practice. To produce research of the highest possible quality, it is important to strive for improvements in methodological soundness and reliability, by making use of the findings of scholars who research statistical methods used by linguists (e.g. Rayson et al. 2004b; see further 3.4). It is also vital to extend existing methods and test out new ones, as I do in this study with Baker's (2011) "locking" concept (in 2.6).

Above, I have briefly defended the basis of the keyness method used in my study, to obtain empirical data on which a comparative analysis of Shakespeare's language style can be based. As implied by the mention of studies such as Mahlberg (2007), in 2.5.1, and Ho (2011) and McIntyre (2010), in 2.5.2, quantitative methods using frequency and keyness are now applied to other linguistic items besides single

words. In the next section I explain briefly why word clusters and semantic domains are the most useful additional dimensions to single words in my study, rather than other possible options such as parts of speech and "key key words".

2.5.4 Other language units which can be investigated using frequency and keyness methods

Studies of electronically-derived recurrent word combinations (see 1.5) such as word clusters, lexical bundles and n-grams have been demonstrated as useful in the investigation of a number of registers of language. Biber et al. (1999:990-1024), Biber, Conrad and Cortes (2003, 2004) and Biber and Barbieri (2007) have been notably influential, as has Sinclair (e.g. 1996, 1991, 2004). This approach has been extended to literary texts in recent years, by, for example, Culpeper (2011); Fischer-Starcke (2009, 2010:108-143); Mahlberg (2007); McIntyre (2010) and Stubbs (2005).

In his corpus study of Joseph Conrad's novel *Heart of Darkness*, Stubbs (2005:13) argues that analysing combinations of words in literary texts reveals useful information about the pragmatic and discoursal functions of dialogue and narrative, because these aspects of language rely on words working together rather than on their own. Dramatic dialogue lends itself to investigation via statistically significant recurrent word combinations, through which pragmatic aspects such as politeness formulae and chunks or fragments of speech acts can be captured and analysed. Culpeper (2011:73) begins to explore the potential of recurrent word combinations in Shakespeare's plays and other EModE drama, as mentioned in 2.5.1, arguing that, based on the evidence of e.g. Stubbs and Barth (2003), "lexical bundles are good discriminators of different styles". In his investigation of the top 10 most frequently-occurring 3-word lexical bundles in Shakespearean and other EModE drama, Culpeper finds that more of the Shakespearean bundles begin with the pronoun *I*, and that the

pragmatic marker *I pray you* is more frequent in Shakespearean drama (2011:72-74). Similarly, in my previous research into key word clusters in Shakespeare's plays, I found *I pray you* to be associated with female dialogue (Demmen 2009:98-99). Including word clusters in the present study enables me to add much more to the findings from these studies, which demonstrate that language styles do vary in formulaic language. I discuss word clusters further in 3.2.

Wmatrix's facilities for generating lists of parts of speech ("POS") and semantic domains in corpora, mentioned in 2.5.2, are aided by two taggers which annotate the text automatically. POS are annotated by CLAWS (Constituent Likelihood Automated Word-tagging System)[15] and semantic domains by USAS[16] (the UCREL[17] Semantic Analysis System). Rayson puts forward the following argument for analysing key POS and key semantic domains in addition to keywords:

> Key grammatical categories and key semantic domains are used to group together lower frequency words and multiword expressions which would, by themselves, not be identified as key, and would otherwise be overlooked. (2008:543)

However, Culpeper (2009:54) tests out and quantifies what value is added by key POS and key domains to a keywords analysis, and he argues that "a straight keyword analysis revealed most of the conclusions". Whilst acknowledging that the POS and domain analyses pick up some potentially interesting results which are not of sufficiently high frequency to occur as keywords, he finds that only a quarter of the POS and a third of the domain results actually add any new outcomes (2009:54). Nevertheless, in his (2011:75-78) study, Culpeper shows that the analysis of words which group together into particular semantic domains has potential value as a dimension by which Shakespeare's plays can be profiled in a corpus-based dictionary

[15] See http://ucrel.lancs.ac.uk/claws/ (last accessed 10.08.12) and e.g. Leech et al. (1994).
[16] See http://ucrel.lancs.ac.uk/usas/ (last accessed 10.08.12) and Rayson et al. (2004a).
[17] UCREL = University Centre for Computer Corpus Research on Language.

or encyclopaedia. Archer et al. (2009), too, show that semantic domain analysis is useful in exploring the concept of love in Shakespeare's plays. Furthermore, as mentioned in 1.1, Crystal (2008:155) argues that by grouping "difficult" words in Shakespeare's plays according to their semantic meaning "we can more clearly see the relationships between them". Categorising the words in both my corpora according to semantic domains is therefore helpful in providing a clearer picture of concepts which are less familiar in the present day, through putting them into larger and more distinctive groups, providing of course that the categorisation technique is reliable (see further 3.3.1). Semantic domain analysis is not without problems, however, particularly when applied to EModE (as is indicated by Archer et al. 2009 and Culpeper 2011), and I discuss these further in 3.3.

My analyses do not include "key key words" and their "associates", Scott's (e.g. 1997, 1999) terms for keywords which are linked through occurring as key in multiple texts. These can be generated using *WordSmith*, but the process requires a large number of separate texts (Scott's 1999:Help menu suggests around 500). That is likely to be a limitation of using this method with corpora of literary texts, many of which are relatively small (as argued in 2.3).

In this section (2.5.4) I have argued that investigating statistically significant word clusters and semantic domains provides additional value to the analysis of single words in my study of language style in EModE drama. In 2.5 overall, I have presented evidence supporting the analysis of those which are of high frequency, and those which are key when Shakespeare's plays are compared to a parallel corpus of other contemporaneous plays. However, as I argue in the next section, because the keyness method is oriented towards finding differences between two corpora, similarities

between Shakespeare's language style and those of his contemporaries needs to be brought back into the picture in order to provide a balanced view.

2.6 From keyness to locking: investigating similarities between corpora

In this section I begin by explaining why looking at the differences between texts, which is what the keyness method is designed to do, needs to be supported by an examination of the similarities (in 2.6.1). Then, in 2.6.2, I mention some possible ways of automating the analysis of similarities, and I explain how I use Baker's (2011) new concept of "locking" as the means for doing so in my study.

2.6.1 Why does similarity matter?

The potential advantages of using the keyness method to help find differences in language styles are clear from the existing research discussed in 2.5. However, as stated in 1.1, Baker points out in his (2004:349) study that the contrasts highlighted by keywords are only part of the picture when two corpora are compared, and that similarities should not be overlooked. Indeed, the essential arguments in favour of investigating lexical patterns in large bodies of text using electronic methods (greater systematicity and objectivity in the identification of stylistically important language features, the ability to find lexical patterns that are not apparent through manual analysis, and the possibility of investigating much larger collections of text than could be done manually; stated in 2.4.2) seem no less applicable to the investigation of similarities in texts than to differences between them.

Nevertheless, in corpus stylistic studies which employ the keyness method to investigate language differences and distinctions, similarities remain very largely unaddressed (Ho's 2011 study, discussed further below, being a notable exception).

This seems strange, since other corpus-based approaches do account for similarity as well as difference. For example, that is how the computational stylisticians Craig (2011), Elliott and Valenza (2011) and Rosso et al. (2009), discussed in 2.4.3, are able to determine that Shakespeare's vocabulary is similar to that of his peers. It is also the way Hope and Witmore (2010:188-192, 387-389) put rhetorical features in Shakespeare's plays into a bigger context, mentioned in 1.1, by showing the relative similarity between each one and nearly 300 other contemporaneous plays (using the empirical data extracted from their corpus; see further 3.3.1). Similarity is also a factor in diachronic research, for example the studies of Baker (2011) and Hilpert and Gries (2008), who measure language change in English over the 20th century. I discuss Baker's (2011) study in more detail below.

The orientation of stylistics to language deviations (from a particular norm), discussed in 2.2 and 2.4.1, perhaps intuitively lends itself to a search for difference rather than similarity, and hence to using a method such as keyness. Certainly, describing how language styles are different from one another leads to a greater understanding of what they are like. However, enlarging this with some comparative discussion of how language styles are also similar provides a more balanced view than viewing differences in complete isolation from similarities. Importantly, it reduces the risk of overstating differences and distinctions (a point made by Baker 2004:349, introduced in 1.1), in the language styles of characters, authors, and indeed text-types, genres and registers. It can also provide some statistically-based information about the comparative norms between two texts or two corpora, as I argued in 1.1. This is also important, because theories of foregrounding (Jakobson 1960; Mukařovský 1964a and b) rest on the notion that some language features stand out as noticeable through deviating from other language norms. As I argued in 1.1, while the keyness method in

corpus linguistics can usefully aid in the investigation of foregrounded language in stylistic studies, it does not actually reveal any information about the norms, which remain unidentified. In a corpus stylistic study it could be argued that the language which deviates the least, statistically, between two texts or corpora (for example, by having the most similar frequency) constitutes a norm of some kind. A method of identifying some language norms would be useful in research which use keyness to help automate the investigation of foregrounding in stylistics (although that is not an aspect of the present study). However, it must be remembered that foregrounding is a psychological notion (as noted in 1.1; see in particular van Peer 1986). This means that the relevant background to what is foregrounded may not necessarily correspond to statistically quantifiable language units (see further 7.5.3).

For the above reasons, I would argue that addressing similarity is an essential component of corpus stylistic research. I would also propose the following reasons for the present lack of attention to similarities:

(i) an assumption that similarities between texts are obvious, or can be discerned intuitively, and therefore are not worth attempting to quantify and analyse; and

(ii) a lack of knowledge and/or availability of automated processes for applying appropriate statistical methods to highlight similarities between corpora, to make it realistically achievable for non-statisticians.

With regard to point (i), Ho (2011) shows that similarities between texts are by no means all intuitively obvious. Her study is a rare exception in corpus stylistic research, since her investigation of language differences between the 1966 and 1977 editions of John Fowles' novel *The Magus*, using key semantic domains and *Wmatrix*, is supported by an examination of empirically-based similarities. This enables her to comment on aspects of the language which Fowles chose **not** to change, as well as

those he did, giving perspective on how far the two editions vary. In my study, discussing Shakespeare's style in the context of its similarity to other EModE plays adds a correspondingly valuable dimension, providing a greater understanding of how far Shakespeare's style is and is not like those of his peers.

Point (ii) above raises the question of how data on statistically-based similarities can be extracted from corpora. Ho (2011:70-83) achieves the analysis of similarities between the two editions of *The Magus* using programmes which are particularly suited to comparing different versions of the same text. These are TESAS[18]/*Crouch*, which quantifies textual re-use/revision and has been used in journalism, and *WCopyfind*, a plagiarism-detecting programme. They are not suitable tools for my study, however, since I am comparing two entirely different sets of texts. Other potential ways of automating the analysis of language similarity between corpora include "consistent collocates", investigated by Gabrielatos and Baker (2008) in their corpus-assisted discourse study of the representation of refugees and asylum seekers in the UK press, and "lockwords" (Baker 2011). I use the latter method in my study, (a) because it offers a useful direct contrast to key results, and (b) in order to take up the opportunity of assessing how well it can be applied using the relatively quick and user-friendly keyness software tools in a slightly different way. I explain this in the next section.

2.6.2 Addressing textual similarities using keyness software tools

Baker (2011) uses several frequency-based statistical methods to determine the words which remain consistently important over time in four diachronic corpora of 20th century British English, from which he puts forward the concept of "lockwords".

[18] TESAS = TExt Source Alignment System.

These are high-frequency words which occur statistically with the most similar frequency when the corpora are compared. Baker explains that a lockword is:

> a word which may change in its meaning or context of usage when we compare a set of diachronic corpora together, yet appears to be relatively static in terms of frequency. (2011:66)

and that lockwords provide a statistical contrast to keywords in corpus linguistics:

> These words were so consistent in their frequencies that they appeared to be the opposite of Scott's (2000) concept of keywords [...] a new term, *lockword*, was thus invented to describe them. The term *lock* was chosen because it is related to *key* (*key* is the highest collocate of *lock* in the British National Corpus (using log likelihood), and furthermore, lock is a good description of these words: they appear to be "locked" in place. (2011:73)

Baker (2011) successfully uses this method to identify words which are locked over the variable of time (for example *money*). The concept of statistical locking does not have to be limited to a diachronic perspective, however. It can equally be applied to synchronic corpora (Baker, personal communication, 25.10.11), through which it shows what the language in the texts have most strongly in common, on an empirical basis. In my study, therefore, it enables me to find out:

- how the language used by Shakespeare in constructing character dialogue is similar to that of his contemporaries; and
- how EModE plays are characterised by shared preferences for some language styles among playwrights

based on data from about 1,600,000 words of dramatic dialogue in 79 plays by 24 different playwrights (taking the two corpora overall; see further chapter 4).

As stated in 1.1, my study also extends Baker's (2011) lockwords method, in three ways:

(i) through its application to synchronic rather than diachronic corpora, as indicated above;

(ii) by testing it with historical texts; and

(iii) in using a different process which orients the statistical computations of the keyness tools in *WordSmith* and *Wmatrix* to similarity rather than difference, with relative simplicity and speed, as I explain below.

Baker (2011) extracts lists of the most frequent words in each of his four corpora using *WordSmith*, and then uses the statistical analysis software *SPSS* to compare them and identify those with the most similar frequencies (using standard deviation and co-efficient of variance[19]). I have only two corpora, however, so I am able to use the keyness tools in *WordSmith* and *Wmatrix* to identify words and other linguistic items which occur statistically with the most similar frequency, based on log-likelihood statistical tests (Baker, Hardie and Wilson, personal communication, 25-27.10.11) (*WordSmith* and *Wmatrix* are currently limited to comparisons of two corpora only). Setting the p value to 1.0 (in *WordSmith*) or the log-likelihood value to 0 (in *Wmatrix*, which does not provide p values), causes the keyness tools to identify words which are the **least** key, i.e. those for which the software finds the least information indicating there is a difference in frequency between the two corpora. Log-likelihood values at or near 0 occur either when frequencies in both corpora are low, or when frequencies are relatively similar (Rayson, personal communication, 19.12.11). Excluding low-frequency items restricts the results to the latter kind, which constitute the locked results (minimum frequency settings are discussed in 3.4.2).

Using specialised corpus linguistic keyness tools in *WordSmith* and *Wmatrix* to identify lockwords offers several advantages over Baker's (2011) method:

- speed and ease;

[19] Standard deviation "measures the spread of data from the mean frequency of a word" and coefficient of variance (standard deviation divided by mean average, then multiplied by 100) controls for frequency (Baker 2011:72).

- access to contextual data through concordances (facilitating investigations of how words with the most similar frequencies are used in each corpus);
- computations which take into account the relative sizes of the corpora (Baker, personal communication, 25.10.11)[20]; and
- a statistical testing method which has been more rigorously tested in corpus linguistics (e.g. in Rayson et al. 2004b); log-likelihood values can also be looked up in tables of statistics, unlike co-efficient of variance values (Rayson, personal communication, 19.12.11).

Baker's (2011:66) initial definition of a lockword, given above, seems to allow for diverse functions in words which are of similarly high frequency, statistically, in corpora. However, later in his study (2011:83) he mentions the distinction of "a true lockword": one which not only has a similarly high frequency in the corpora being examined, but which is also used in the same way(s). This raises the important question of whether or not all words with similarly high frequency which are matched orthographically by the computer software can be considered "locked", or whether there should be an additional linguistic criterion based on similarity of function.

A parallel question can of course be applied to keywords and other key results: does everything on the list of output generated by the software really count as a keyword, or only those items which are useful to the researcher? Based on Scott's (2010:46) comments regarding keywords and keyness, noted in 2.5.3 above, it would be difficult to get agreement on what should count as a "true" lockword, since researchers would undoubtedly make different judgments about similarity of function and/or other qualifying features. It seems better to set the criteria for locking and keyness according to statistical parameters only, then to assess all the results in terms

[20] Although my corpora are of very similar size, they are not identical (see 4.4).

of prototypicality and usefulness for the researcher's purposes. Key or locked results which are problematic can be diagnosed, and if necessary disqualified from further analysis, with the reasons for this made clear in relation to the aims of the study. This is a practical approach with a view to application, since my ultimate goal in obtaining locked and key results is to produce empirical data that will most usefully launch detailed investigations of potential style features in my corpora.

Having explained the principles of locking, and argued that it can usefully be applied in my study, I end this chapter with some further discussion of the kinds of results I anticipated from my corpus data, and my approach to interpreting them. This is based mainly on theory and background from keyness studies, but the principles extend equally to results generated with the locking method.

2.7 Issues surrounding the interpretation of corpus results in stylistic analysis

Other scholars working with keywords note several types which typically occur in any given output (see for example Culpeper 2009:38-39; Scott e.g. 2000; Scott and Tribble 2006). These are:

- proper nouns;
- results which reflect the "aboutness" of the text(s) in the corpus. Although they may be topical, they are not limited to topic or theme. As McIntyre (2010:169) explains, "aboutness" keywords "reflect/create particular characteristics of the text's genre" (and see further Culpeper 2009:38-39);
- results which are evidence of style features.

Culpeper (2009:35) also distinguishes between "generalised" and "localised" key results, depending on their frequency and distribution in the texts in a corpus (discussed further in 3.4 and 3.5).

The abovementioned types of results were reasonable to anticipate finding in the output of my keyness investigations (in chapter 8), but since there are as yet no studies of locked results in EModE or in drama, it was much harder to anticipate what kinds of words, clusters or semantic domains would arise in the locked output from my two corpora. The prospects were therefore quite exciting. I expected to find some evidence of "aboutness" in the locked results, but that it would be more generalised than that in the key results (for example, reflecting broad themes which were popular in English drama in the late 16th and early 17th centuries). That expectation was broadly realised, as discussed fully in chapter 7, where I also note the occurrence of a proper noun among the lockwords. This is more surprising, as proper nouns tend to be relatively localised and topical, and therefore seem more likely to occur as a marker of language difference (i.e. in key results), not similarity between corpora.

I also expected that, like key results, not all the locked results would necessarily lead to stylistic insights. This is because computer-generated data from literary texts contain potential candidates for style features, not guaranteed candidates, as indicated in my discussions in 2.5.3 above. As Leech (2008:164) puts it, the output consists of "a set of linguistic features which are empirically derived 'good bets' to follow up in undertaking a subsequent stylistic analysis". Other corpus stylisticians make the same argument in slightly different terms: Archer and Bousfield (2010:203) state that "keyness analyses are at best a 'way in' to a text"; Culpeper (2009:32) stresses that "the link between keywords and style" is what is important, rather than the fact that some items stand out statistically from others, and McIntyre (2010:168) emphasises that "statistical significance does not necessarily equate to interpretative significance". Furthermore, with regard to stylistic analysis in general (corpus-based and otherwise), Leech (2008:24) emphasises that not all deviations from any given

norm of language have a stylistic effect; he argues that some are "unmotivated deviations which have a trivial and unintended meaning". Therefore, whilst it is important to maintain a systematic approach to avoid accusations of circularity such as those noted in 2.4.2, it is also crucial to be discerning over results which will most usefully benefit from close qualitative analysis, and thereby further the researcher's aims. Categorising results in various ways can help with the identification of those which are most useful, as I discuss next.

Scott and Tribble (2006:63) indicate that topical keywords are less useful for stylistic analysis. However, other research indicates that the inclusion or exclusion of "aboutness" results, proper nouns and/or results which are topical or localised depends on the purposes of individual researchers. For example, McIntyre (2010:169) opts for "filtering out those proper nouns and aboutness keywords which have no bearing on characterization, and focusing on those that we might be less inclined to predict through intuition". Culpeper (2009:39), on the other hand, takes a more inclusive approach by categorising all his keywords according to Halliday's (e.g. 1994) "metafunctions" of language to gain an overall view of the proportion which have interpersonal, textual and ideational functions in the play. Then, he discusses in more detail the keywords which have the greatest implications for characterisation. Mahlberg (2007) focuses particularly on key word clusters with localised functions to investigate characterisation in Dickens' novels, whereas in order to investigate the language styles of male and female characters in Shakespeare's plays (Demmen 2009), I focus mainly on those with generalised functions.

Regardless of differing analytical approaches and criteria, the above scholars concur that careful qualitative research is vital in order to get value from the keyness method. The fact that keywords (and, by extension, lockwords) fall into identifiable

types leads to the question of whether it is desirable to categorise quantitative results in a formal way, as a precursor to qualitative analysis. Other corpus stylistic studies use a range of approaches from frameworks with a theoretical basis, to intuition-based categories, to informal grouping of results during the course of discussion. At the more theoretical end, Culpeper (2009:39) argues that analysing keywords from a functional perspective allows for better access to the potential pragmatic and discoursal effects of words which surface as key. Like Culpeper, Leech (2008:136-161) uses Halliday's metafunctions to classify keywords in his analysis of Woolf's *The Mark on the Wall* (a short story in the form of an "interior monologue"), though he concludes that an overall "expressive" function is a more satisfactory way of describing language which represents the internal thoughts of one narrator. This is less of a problem with drama, since the dialogue is written for an on-stage rendition of the fictional world. Interactional dramatic dialogue, being an approximation of naturally-occurring speech, lends itself better to ideational, interpersonal and textual descriptions of function, as discussed further in 3.2.5 (although soliloquies arguably have more of an expressive function).

Fischer-Starcke (2009) and Mahlberg and Smith (2009:452-453) use intuition-based categories to group their keywords from Austen's *Pride and Prejudice* in a convenient way, e.g. "family relationships" (Fischer-Starcke 2009:501-508) and "civility" (Mahlberg and Smith 2009:453). B. Walker (2012:98) adopts a range of approaches to give different perspectives on his key results from Barnes' novel, including a result-by-result analysis of keywords, some "ad hoc grouping" and a formal framework of analysis (Culpeper's 2001 textual cues to characterisation).

From the abovementioned studies, I would contend that the benefit of classifying key results depends very much on the aims of the research, the number of

results to be dealt with, and the ways in which individual researchers find it most helpful to think about their results in the course of analysing them. In my study I am necessarily limited to analysing fairly small numbers of results of each type, to accommodate discussions of the three methods I am using (frequency, keyness and locking), applied to three different types of language unit (words, word clusters and semantic domains). I do not seek to link the output to any particular theoretical basis, but to a wide range of other linguistic and literary studies which are relevant to EModE drama, as stated in 1.2. Applying too many pre-determined categories would unduly constrain this. However, it is useful to the analysis of my word cluster data (in 6.3, 7.3 and 8.3) to include a comparative discussion of Culpeper and Kytö's (2010) findings from lexical bundles in EModE drama and other speech-related registers, since my data is very similar to theirs but from a different source. Their data is classified according to a framework of functional categories which are derived from EModE speech-related texts, and therefore data-driven (see 2.3). I explain this framework and its application in more detail in 3.2.5.

My semantic domain data is, like that in other studies such as Archer et al. (2009), Ho (2011) and B. Walker (2010, 2012), classified automatically by *Wmatrix*, as explained in 3.3.1. I could see no analytical advantage to classifying the keyword results (or single high-frequency words or lockwords) into categories in this study, and I simply discuss any obvious groupings, making clear why the ones I analyse in more detail have greater potential bearing on language styles in EModE drama. To assist in this, I use the concordance data for keywords and other results from *WordSmith* and *Wmatrix*, to see them in context and assess what functions and effects they have. Concordances enable the user to pinpoint the location of every instance of a result in the corpus texts, from which it is possible to see the kind of speech act it occurs in, and

in what circumstances (for more about this function, see Scott and Tribble 2006:33-53). This is vital to understanding pragmatic effects in dramatic dialogue.

2.8 Summary

In this chapter I have explained the methodological principles which underlie the analysis of the empirical data I use in my investigations of language style in plays by Shakespeare and other contemporaneous dramatists. I have discussed the relationship between corpus linguistics and stylistics (in 2.2 to 2.4), and how the former can serve the aims of the latter by yielding up quantitative data which points to prospective language style features. I have argued (in 2.5 and 2.6) that, from a range of statistical methods which are possible using *WordSmith* and *Wmatrix*, simple frequency, keyness and the new method of locking are the most useful ways of finding potential style features in the dialogue of the plays. I have also justified the benefits of applying the methods to word clusters and semantic domains as well as to single words, in order to find more evidence of how Shakespeare's style both is and is not like that of a range of his peers.

During the course of my discussions I have charted existing corpus-based research into Shakespeare's plays, and I have argued, in particular, that my study addresses a current lack of corpus stylistic research which compares them to other EModE drama. I pointed out in 2.5.1 that most existing corpus stylistic research into Shakespeare's plays is limited to internal variation (with or between Shakespeare's plays), but that Culpeper's (2011) initial findings (mentioned in 2.4) indicate that a larger comparative study with other contemporaneous drama would be rewarding.

I have also argued that my study helps address a lack of attention to similarities in language style in existing corpus stylistic research, in 2.6. There, I also made the

case for adjusting the keyness tools in *WordSmith* and *Wmatrix* to apply Baker's (2011) "locking" concept, to identify language which occurs with relatively similar high frequency in my corpora of Shakespeare's plays and other contemporaneous plays.

In the next chapter, I discuss some specific issues surrounding the investigation of word clusters and semantic domains in more detail, as well as some operational matters which have bearing on the successful application of the corpus methods detailed in this chapter.

CHAPTER 3. WORD CLUSTERS, SEMANTIC DOMAINS, AND OPERATIONAL CONSIDERATIONS

3.1 Introduction

This chapter, like chapter 2, answers research question 2 concerning corpus linguistic methods. I provide further detail about the linguistic items that constitute word clusters and semantic domains in 3.2 and 3.3, respectively, and how they add useful analytical dimensions to my investigation of language styles in EModE drama. I also highlight some issues and problems which can be anticipated from existing research. In 3.4 I discuss the operational settings and parameters used for the corpus linguistic software tools in the study, and the important bearing that these have on the empirical data, on which my findings are based. I consider the distribution of results in corpora in 3.5, and the influence and implications of reference corpora in 3.6.

3.2 Word clusters and other types of recurrent word combinations

In 3.2.1 I summarise the theoretical background to the investigation of recurrent word combinations such as word clusters (terms which were defined in 1.5), before explaining the way they can be captured with corpus linguistic methods. This includes further clarification of the particular concept of a "word cluster" in my study. Next, in 3.2.2, I explain why word clusters are a helpful contribution to an investigation of language styles in plays by Shakespeare and other contemporaneous dramatists. In 3.2.3 I discuss the word cluster length used in this study, and in 3.2.4 I explain the functional categories which I apply to word clusters in my analyses (in 6.3, 7.3 and 8.3).

3.2.1 Investigating recurrent word combinations with corpus linguistic methods

Studies of various types of recurrent word combinations via corpus linguistic methods have become increasingly popular, and have been carried out in a range of registers and text-types. This is discussed in more detail by, for example, Biber, Conrad and Cortes (2003, 2004:371-372) and Culpeper and Kytö (2010:104-105). The study of recurrent word combinations is based upon arguments that language is to some extent "formulaic", i.e. that it is stored in the mind in pre-prepared sequences or "chunks" (see e.g. Schmitt and Carter 2004; Wray 2002, 2008, 2009). Culpeper and Kytö (2010:131, 140) find many lexical bundles in their speech-based EModE data which fit Biber et al.'s (1999:1073) concept of "utterance launchers" (having a subject pronoun + verb phrase construction, and beginning an utterance, e.g. *I would not*). These provide a formulaic start, which give speakers a chance to prepare the remainder of what they intend to communicate (Culpeper and Kytö 2010:93-94, 107, 140). This efficiency measure from natural speech is also built into dialogue constructed by dramatists, and I highlight the presence of utterance launchers in my data in 6.3.

Research into PDE shows that language formulae are not merely time-saving lexico-grammatical constructions: they are also "primed" in the speaker's mind with information which has social implications (Aijmer 1996:8; see further Goffman 1971; Ferguson 1981; Hoey 2005, 2007; Morley and Partington 2009:145-146). This makes language formulae which are used relatively frequently by characters in dramatic dialogue well worth investigating, because they provide access to politeness routines and other ways that Shakespeare and his peers construct characters as engaging in social relationships, over and above what is revealed by quantitatively significant single words. Although dramatic dialogue is scripted and embellished in ways that natural speech is not (Short 1996:174-179), dramatists necessarily call upon their own

primed knowledge of conversational routines and ritual language formulae in constructing social situations and relationships with which their audiences are familiar. As Short (1996:179) points out, for an audience to relate to a play there has to be "communicative correspondence between ordinary conversation and drama".

Though language formulae are argued as being mental as well as lexical phenomena (Morley and Partington 2009:145), the fact that they surface at the lexical (or lexico-grammatical) level is what enables them to be investigated through the frequency-based recurrent word combinations generated by corpus linguistic software such as *WordSmith*. Stubbs argues that:

> Phraseology and corpus study are intimately related, because it is scholars who have studied texts and corpora who have been struck by the large amount of recurrent phraseology which characterizes normal language use. (2007:163)

The computational process of locating and extracting recurrent word combinations with corpus tools such as *WordSmith* is summarised by Culpeper, as follows:

> Essentially, the computer works through the text, recording the co-occurrence of every word with its neighbours, and then calculates which groups of words most frequently co-occur. Multiword units, thus defined, may be considered a kind of extended collocational unit and are frequently referred to as lexical bundles or clusters. (2011:72)

As indicated in 1.1, collocations are words which frequently occur near one another (see, e.g., Biber et al. 1999:988-990; Hori 2004; Scott e.g. 1999:Help menu). Computer-identified recurrent word combinations do not necessarily reflect formulaic chunks stored in the mind of any particular human language user, however, even though the word combinations may be recognisable and familiar. Schmitt and Carter (2004:2) note that although corpus linguistic studies can potentially help in tracing formulaic language, frequency of a recurrent sequence of words in a text, found by a computer, may not necessarily equate to a formula in the mind of the speaker (or

writer). As Scott and Tribble (2006:41) point out, the word cluster output produced by *WordSmith* constitutes "simply the repeated strings found most often" in the texts.

As indicated in 1.5, there are different terms in use for recurrent word combinations, since they are conceptualised and described in various ways by scholars focusing on different language aspects. There are, however, issues of compatibility. Here, I discuss only the types which are relevant to my research (but see further Culpeper and Kytö 2010:104-105). Biber et al. (1999:990-1024) and Biber, Conrad and Cortes (2003, 2004) use the term "lexical bundle", a concept which is taken up in the investigation of EModE by Culpeper and Kytö (2010) and Culpeper (2011:72-74). Scott and Tribble (2006) use the term "word cluster", which is also used by Mahlberg (2007) in her research into Dickens' novels. Scott and Tribble (2006:12, 32:note 1) argue that lexical bundles and word clusters are essentially the same concept. However, although both are generated using the word cluster facility in *WordSmith*, it is worth noting that not all the *WordSmith* word cluster output necessarily meets Biber et al.'s criterion of being non-local or non-topical, below, which applies to lexical bundles:

> Local repetitions typically reflect the immediate topical concerns of the discourse. In contrast, lexical bundles can be regarded as lexical building blocks that tend to be used frequently by different speakers in different situations. (1999:991)

Therefore, I would contend that in principle a lexical bundle is a non-localised form of word cluster, though in practice (i.e. in corpus output) they may well be the same. I do not formally exclude localised or topical results from my discussions since they can be of potential interest, though I do minimise them (by adjusting the parameters of the software; see 3.4.2-3) in order to prioritise results which characterise all or most of each corpus. These are more likely to point to authorial style features in Shakespearean and other contemporaneous dramatic dialogue. Therefore, in principle

it is more accurate to describe my recurrent word combination data as "word clusters", though in practice most of them also qualify as "lexical bundles".

3.2.2 The benefit of including recurrent word combinations in an investigation of language style in Shakespeare's plays and other contemporaneous plays

Studies of prose fiction show that insights into the language styles of authors of well-known literary texts can be gained by analysing recurrent word combinations. For example, Hori (2004) and Mahlberg (2007) both focus upon the novels of Charles Dickens, Hori examining collocations and Mahlberg investigating word clusters. Fischer-Starcke (2009, 2010:108-143) investigates frequent phrases in Jane Austen's novels. As discussed in 2.5.1 and 2.5.4, Culpeper (2011:73) argues that lexical bundles are a useful route to profiling language styles, in his study of Shakespearean and other EModE drama. My own previous research also shows that recurrent word combinations lead to information about the construction of dramatic genres (Demmen 2007) and the language styles of male and female characters (Demmen 2009).

Relatively little research into recurrent word combinations in EModE drama has thus far been done, although the arguments of Craig (2012) and Crystal (2008), who have both conducted substantial corpus studies of Shakespeare's plays, imply its potential value for unlocking some of the mysteries of his language style, as mentioned in 1.1. Craig (2012:6), whose research indicates that Shakespeare's vocabulary is not exceptional compared to that of his contemporaries, suggests that "[i]f Shakespeare is not unusual in his word frequencies, is there something about the larger units of his language that is more remarkable statistically?" Crystal (2008:173) states that collocations are underexplored in Shakespeare's plays, claiming that collocations "by their nature present us with more striking images than we find in individual words [...] It is these juxtapositions of images which stay with us, and

which provide us with much of Shakespeare's quotability." Including word clusters in my corpus stylistic investigation of Shakespeare's language style in the context of that of his peers therefore helps to address an under-researched area, in addition to adding a useful analytical dimension to my study.

Analysing word clusters also enables me to make some comparisons with Culpeper and Kytö's (2010:103-141) findings from lexical bundles in EModE drama, since our data is similar (as indicated in 3.2.1). Their research is focused on variation between EModE speech-related text-types, rather than on authorial style, as mine is, and they use a different corpus (the *CED*, mentioned in 2.3 and discussed further in 4.3.1). Some comparisons therefore usefully add to what is known about the nature of formulaic language in the register of EModE drama. Moreover, they provide an opportunity to verify that my new and untested corpus of EModE plays contains similar language features to those in another well-researched corpus of the same text-type, which would be reasonable to anticipate.

3.2.3 Word cluster length

The cluster facility in *WordSmith*, mentioned in 2.5.1, can identify repeated strings of between two to eight words[21]. Tests with my corpora showed that 2-word clusters are numerous, but contain little of the pragmatic or discoursal information which I wanted to capture for my analyses. As argued in 2.5.4, this is important to the investigation of style in interactional dialogue. On the other hand, clusters longer than three words were too few to constitute useful sets of results. Accordingly, I confine my analyses to 3-word clusters, noting any which overlap into longer formulae where relevant, in my discussions in chapters 6 to 8. Focusing on 3-word clusters also maximises the

[21] At the time of writing, the cluster/n-gram facility in *Wmatrix* is unavailable but under development. Other software exists, for example Fletcher's (2002-2007) *kfNgram* programme, used by Fischer-Starcke (2009, 2010).

potential for comparative discussions of Culpeper and Kytö's (2010:103-141) lexical bundle data, mentioned in 3.2.2 and in 2.5.1. As indicated in 2.7, to facilitate closer comparison I categorise my word cluster results according to the framework of functional categories they use, which I detail next.

3.2.4 Functions of word clusters in Early Modern English plays

Although other functional frameworks have been used for recurrent word combinations, these are less suitable for EModE plays than that of Culpeper and Kytö (2010:107-134), either because they are derived from language constructs that are different to word clusters, and/or through being designed for other text-types. For example, Moon's (1998) functional categories are designed for PDE (mostly newspaper text), and for fixed expressions and idioms, which are more restricted categories of recurrent word combinations than the word clusters in my data. Mahlberg's (2007) functional categories for word clusters in the novels of Charles Dickens, mentioned in 2.7, are oriented towards localised results, which are not the main focus of my analyses. Culpeper and Kytö's (2010:107-134) framework is specifically designed for data similar to the word clusters in my study, and also for EModE, including drama. I applied their framework in my previous research into recurrent word combinations in Shakespeare's plays. This enabled me to show, for example, that women in Shakespeare's plays talk relatively more than men about teaching and learning, and that women make relatively more use of negative volition as a language strategy (2009:174). Accordingly, I now explain the basis of the functional categories.

Various theoretical frameworks of language function can be applied in stylistic analysis (see e.g. Leech 2008:104-117), but the only one I discuss is that chosen for

my own study. Its main categories are Halliday's "metafunctions" of language, which have been used in the analysis of Shakespeare's plays by B. Busse (2006) and Culpeper (2009) as well as in Culpeper and Kytö's (2010) study of speech-related EModE. These overarching functional categories are summarised below (based on Halliday e.g. 1994:179-180):

- **Interpersonal** functions establish relations between speaker and addressee, for example by conveying information about the speaker's mood (through modals) or about relationships between speakers and addressees (through the kinds of speech acts they use to one another).
- **Textual** functions organise the message of the text in a meaningful way, for example the manner in which the information is ordered.
- **Ideational functions** make reference to something, for example states, events or ideas. States include literal and metaphorical states.

Busse (2006:e.g. 57-61) uses Hallidayan metafunctions in her study of vocatives in Shakespeare's plays, and goes into further detail of Halliday's (e.g. 1978:48-49) "logical" and "experiential" sub-categories of the ideational metafunction, and other aspects of Systemic Functional Grammar. However, these are not part of Culpeper and Kytö's (2010) approach, and so are not utilised in my study. Culpeper and Kytö (2010:110-111) create sub-functions of Interpersonal, Textual and Ideational categories which incorporate the "macro-categories" of lexical bundle functions proposed by Biber, Conrad and Cortes (2003). Below these, Culpeper and Kytö derive a third hierarchical layer of individual functions based on the kinds of results they find in their data (taking into account the co-text and context of those results).

The categories in Culpeper and Kytö's (2010) framework accommodate several registers of EModE, but some of the register-specific functions in their data do not

arise in my dramatic dialogue data. I therefore use a version of their framework which is streamlined to exclude categories for which I have no data, which I also used in my previous research (see Demmen 2009:74-85). The functional categories used in this study are shown in Table 1 on the next page, with some explanatory notes in italics.

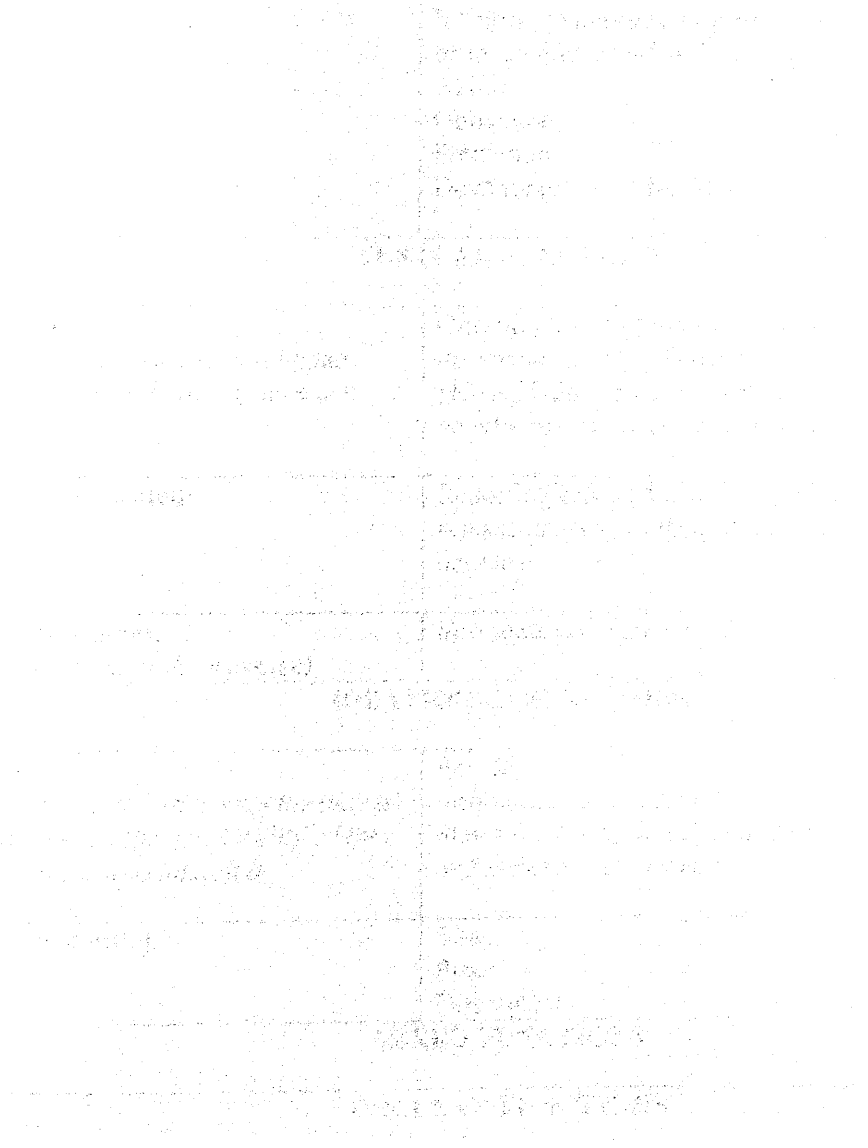


Table 1. Functional classifications of word clusters in this study (based on Demmen 2009:81; categories adapted from Culpeper and Kytö 2010:103-141)

INTERPERSONAL FUNCTIONS	
Speech act-related: *(carrying some force or purpose beyond the words themselves)*	Directive *(having the purpose of telling or asking another character to do something, however strongly or weakly; includes e.g. requests, commands)* Assertion Expressive Sincerity device Vocative Thanking
Modalizing: *(showing degrees of a character's attitude or opinion)*	Volition *(a character's desire for an outcome, or negated: a desire for something **not** to happen)* Intention *(a character's plans to do/not do something or to achieve/avoid an outcome)* Ability Obligation Prediction Downtoners/amplifiers/hedges/emphatics
TEXTUAL FUNCTIONS	
Discoursal: *(a communicative act at a higher discourse level than a speech act)*	Question *(in this framework, a question is a discourse act, based on Sinclair and Coulthard's (1975:27-28) concept of an act which elicits, directs or informs, and not a speech act)*
Narrative-related:	Reporting/reported clause fragments *(includes reports of speech, thought, writing and events or actions)*
Organisational: *(arrangement of the message)*	Informational elaboration
IDEATIONAL FUNCTIONS	
Topical: *arising directly from what the play is about, i.e. its themes, locality or the individuals who inhabit it*	People Informational specificity States *(including physical or attitudinal states, literal and metaphorical states)*
Circumstantial:	Time Place Directional
MIXED FUNCTIONS	
UNCLEAR FUNCTIONS	

The categories in Table 1 accommodate word clusters operating on different language levels, which means that all the data can be accounted for, including local or topical results (e.g. in the Ideational: Topical category), some of which were anticipated in my data (as mentioned in 2.7). I assign categories by analysing the function of each word cluster result using the concordance information. In my discussions, I use initial capitals to denote functional category labels (e.g. Question, Directive), to distinguish them from linguistic terms. Functional category labels are hierarchical, and shown as metafunction: category: sub-category (e.g. Ideational: Speech-act related: Vocative). The classification of results is not always straightforward, however, as I will explain.

In their discussions of stylistic analysis using functional approaches, Jeffries and McIntyre (2010:71-99) point out that the notion of "functional categories" and Hallidayan concepts are not without some problems of application, especially to literary language (as also found by Leech 2008:136-161, noted in 2.7). Jeffries and McIntyre argue, for example, that modality can have an ideational function as well as an interpersonal function (2010:77), and the "multi-layered" nature of language is also noted by Leech (2008:106). Culpeper and Kytö (2010:111) acknowledge that this presents a potential difficulty in deciding how the function of a result should be classified, and in my analyses I adopt their solution of classifying each result according to its over-riding function in the message that is being conveyed by the speaking character. That includes the way the addressee responds.

As Culpeper and Kytö (2010:111) point out, this kind of choice is somewhat interpretative. Though I strive to replicate their methods in order to make comparisons between our data, it may be that they would reach different conclusions about the functions of some of my results. For parity, I also follow their method of prioritising functions in the hierarchy of: "discourse act > speech act > lexical/grammatical item"

(2010:111), and I adopt their principle of applying a functional category only where at least half of the occurrences of a word cluster have a particular function. Any results which have a mixed range of functions, none of which account for at least half the cases, are assigned to a "Mixed" category. There is also, in principle, an "Unclear" category to accommodate results which cannot be classified satisfactorily (again following Culpeper and Kytö's approach), though none occur in my data in this study.

I now look more closely at the third type of item in my data: semantic domains.

3.3 Issues in semantic domain analysis

I explained the justification for examining semantic domains as part of my investigation of language styles in EModE plays in 2.5.4. In this section, I discuss some important considerations and limitations of the method, particularly in its application to historical texts. I explain the automatic categorisation of semantic concepts in 3.3.1, and some features of *Wmatrix* which need to be borne in mind in applying the method in 3.3.2. Finally, I look briefly at the potential role of semantic domain data in the analysis of metaphor, in 3.3.3.

3.3.1 Categorisation of semantic domains

The *Wmatrix* USAS tool is pre-programmed with defined semantic categories, into which it allocates words using a system of dictionaries, including one specifically adapted for EModE (see Archer et al. 2003). The main semantic fields into which the USAS tool groups results are given in Table 2 on the next page, and the full tagset of sub-categories is provided in Appendix I.

Table 2. Main semantic categories classified by USAS

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

In my results and analyses (in 6.4, 7.4 and 8.4), the initial letter of the domain tag name corresponds to one of the main areas in Table 2; for example, A3+ ("Being") is a sub-category of "General and abstract terms".

There are some potential difficulties with this kind of automated semantic analysis. In evaluating the benefits of considering the "sense" in which words are used, and not just the frequency with which they occur in corpora, Baker (2004:353) argues that "one problem with combining words into conceptual groups is that it is a subjective process". This is of course further complicated in the study of historical corpora by changes in word meaning which occur over time. Crystal (2008:234-244), for example, mentions a number of "false friends" in Shakespeare's plays, i.e. words which remain in use but whose meaning is now different in PDE (including *awful*, *ecstasy*, *merely*, *naughty* and *rude*, amongst others). The semantic domain categories in Table 2 are based on present-day conceptions of the world which, as argued in 2.2,

are not entirely the same as those of the Early Modern period (though it is difficult to say in exactly what ways they are different). The EModE dictionary does not "know" all the Early Modern senses of the words to which it assigns categories; this is done on a lexical basis, and inevitably the EModE language is effectively processed through a present-day filter. However, the results of the categorisation can be checked, as I discuss in the next section and further in my analyses in chapters 6-8.

Furthermore, it is difficult for scholars to escape their present-day frame of reference entirely in carrying out historical research, however carefully this is done. Hope and Witmore (2010:361) raise similar issues with regard to the automated analysis of rhetorical features in Shakespeare's plays, mentioned in 2.6.1 (using *Docuscope*, which works not unlike the USAS tool in assigning pre-set categories then counting them using log-likelihood testing[22]). They acknowledge that the classification process is based on certain "linguistic, rhetorical, and cultural assumptions" (2010:361), "human interpretations and definitions based on a particular theory of how language works" (2010:365), and that functional classifications cannot satisfactorily take in the multi-functional nature of language[23]. Nevertheless, Hope and Witmore (2010:365) also stress that the inherent subjectivity in the analysis framework is applied in a systematic and consistent way, and they identify language patterns of a complexity far beyond those which could be found through manual analysis, adding to what is known about Shakespeare's style. This is also the case with automated semantic domain analysis, as the few existing studies which apply it to Shakespeare's plays show (Archer et al. 2009 and Culpeper 2011, mentioned in 2.5.4).

[22] See Hope and Witmore (2004); see also http://www.cmu.edu/hss/english/research/docuscope.html (accessed 20.02.12).
[23] The roots of Hope and Witmore's (2010:365) rhetorical framework are Hallidayan.

Since semantic domain analysis is still relatively new, and particularly its application to EModE, using it in my study helps assess the usefulness of the USAS tool further, as well as adding another facet to the investigation of language styles in EModE plays. In my analyses, I point out and discuss a few problems with the semantic categorisation process. With present-day data, the tool is reported by Rayson et al. (2004b) as being 91% accurate, but there are no corresponding figures for EModE. Archer et al. (2009:144) and Culpeper (2009:47) find that, regardless of the dedicated EModE tagger, some of their results are still put into inappropriate semantic domains, and these have to be redistributed manually. Culpeper (2009:47) gives the example of *cousin* being classified by USAS as a person's relation (the present-day semantic meaning) rather than as a friend (the EModE semantic meaning), and he concludes that the USAS tool needs further refinement for detailed study of EModE (2009:79). The EModE dictionary could be extended and enlarged, for example.

Manual examination and reclassification of all the words in a domain category would be time consuming, but clearly necessary for fine-grained analysis of particular semantic concepts. Spelling regularisation improves the reliability of the EModE tagger in the USAS tool (shown by Archer et al. 2003), and I carry this out with my data (as discussed in 5.4). This enables the semantic domain output to provide a general overview of the most frequently-used semantic concepts in Shakespearean and other contemporaneous plays, notwithstanding some mis-tagging. I check the lists of words in each of the domains which occur in my results, and as long as more than half of them are correctly categorised I include them in my analyses.

3.3.2 Features of *Wmatrix* requiring special consideration

It is worth pointing out that the annotation in the *SDC* is not fully compatible with *Wmatrix*, since it is a mix of XML (eXtensible Markup Language) and non-XML tagging (annotation is discussed fully in 5.2). The tags originate in Mike Scott's Shakespeare corpus, on which the *SDC* is based (see further 4.2.2), and they are fully compatible with *WordSmith*. *Wmatrix* requires well-formed XML tags, otherwise it cannot distinguish between the start and end markers of annotation. The *NDC* is annotated with well-formed XML tags which are compatible with *Wmatrix*. However, as noted in my analysis of key results in 8.2, *WordSmith* did not successfully exclude all the contents of the XML speaker-identification tags. For reasons I could not determine (even with expert help), it picked up the word *who* from within the pairs of angle brackets and counted it along with occurrences in the dialogic text. As it was an isolated anomaly, the results are not seriously affected, but this issue does illustrate the difficulties of using annotated corpora with more than one type of corpus linguistic software. The problems *Wmatrix* experienced reading the tagged *SDC* files were more serious, since the programme could process less than half the text in the corpus. Therefore, I created untagged versions of both corpora (for consistency) to use with *Wmatrix* (by globally deleting the XML tags in *Notepad++*). I also joined the individual play-text files into one large file, since *Wmatrix* cannot process multiple files (unlike *WordSmith*).

A further important point (highlighted by B. Walker 2012:106-108) is that the *Wmatrix* concordance data for a word in a particular USAS category displays not only the words with the semantic tag for that category, but all other instances of the word with other tags as well. This muddies the waters considerably when assessing whether or not the categorisation is reliable. For example, there are over 5,000 instances of the

word form *will* in my two corpora. When trying to assess whether or not the 282 instances which are tagged as a personal name (Z1) are correct, I found that the *Wmatrix* concordancer did not isolate them from the other cases of *will* which are verbs marking the future (tagged T.1.1.3) and/or volition (tagged X7+). Walker (2012:106-108) addresses this by exporting the USAS-tagged files from *Wmatrix* into text files and obtaining concordances for word forms with specific tags in another corpus linguistic programme, *AntConc* (Anthony e.g. 2007; hereafter "*AntConc*"). I follow him in doing so for the most frequently-occurring words in each semantic domain in the output from USAS, in order to check that more than half are correctly categorised and to discuss the problematic categories in my data in chapters 6 to 8. I do not carry this out for all the lower-frequency words in each semantic domain, however, because it is very time consuming. An example is shown in the screenshot from *AntConc* in Figure 3 on the next page, showing part of the concordance data for instances of *will* with the semantic tag of X7+ (the domain "Wanted", i.e. having a meaning of volition).

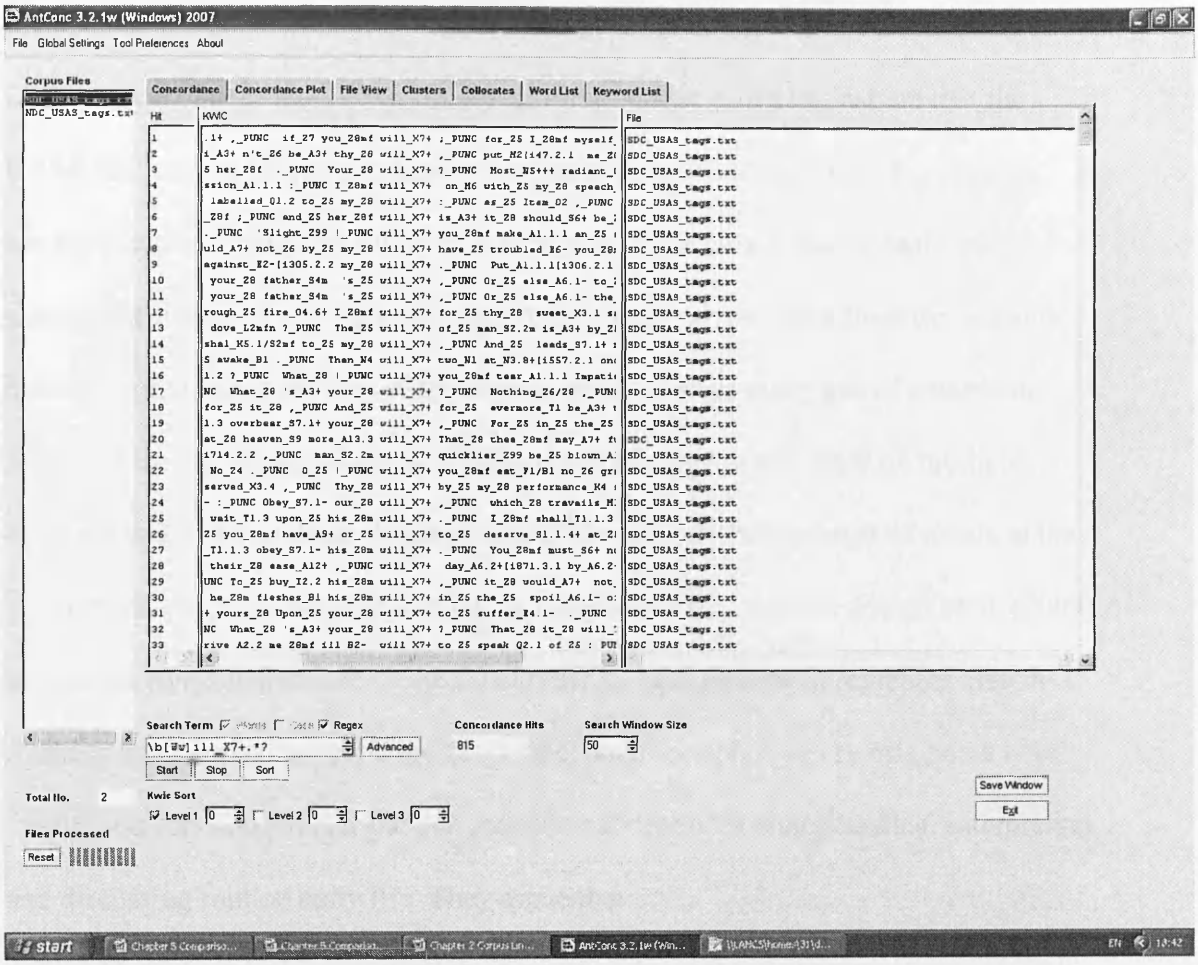


Figure 3. Screenshot of concordance data from *AntConc* for *will:* semantic tag X7+

Since *Wmatrix* does not have a facility for displaying the distribution of results, I also adopt Walker's (2012) method of using *AntConc* to obtain dispersion plots for analysing the distribution of semantic domain results across the corpora. This is achieved by exporting the USAS tagged files from *Wmatrix* into text files, then searching for specific tag labels in *AntConc* and using the Concordance Plot function (Walker 2012:105-108).

3.3.3 Semantic domains and the investigation of metaphor

I anticipated that some of my results would contain figurative language such as metaphor, from the evidence in existing research into Shakespeare's plays by Archer et al. (2009). They show that semantic domain analysis can act as a useful initial filter of

the most common metaphors in a text or text-type, as do studies of PDE texts by Ho (2011) and Koller et al. (2008). However, these studies also demonstrate that the USAS tool cannot distinguish between metaphorical and literal uses of concepts; having classified them semantically, it then groups them on a lexical basis. The metaphorical uses of concepts still need to be identified manually from the semantic domain output, since the tool does not necessarily capture every use of a particular metaphor in one semantic category (because one metaphor can draw on multiple concepts and different lexical items). The comparatively broad range of words in the key domain output (compared to those in a keywords list) enables Archer et al. (2009) to observe some distinctions in the conceptual metaphors used to represent love in Shakespeare's comedies and tragedies. Conceptual metaphor theory originates with Lakoff and Johnson, who argue that metaphor is central to understanding, interpreting and discussing routine daily life. They argue that:

> Metaphors [...] are conceptual in nature. They are among our principle vehicles for understanding. And they play a central role in the construction of social and political reality. (1980:159)

Lakoff and Johnson (1980:159) argue that many conceptual metaphors are longstanding, and a recent and concise summary of conceptual metaphor theory in the context of EModE is given by Oncins-Martínez (2011).

The study of metaphor in historical texts is by no means straightforward or easy, however. In his discussions of sexual metaphors in EModE, Oncins-Martínez (2006) points out that meaning does not remain static over time. More widely, Crystal and Crystal (2002:viii) identify over 47,000 words in Shakespeare's plays likely to present understandability problems today (to the non-specialist in EModE), either through a change in meaning or other usage. They emphasise, though, that this means only about 5-10% of Shakespearean dialogue is potentially problematic for present-

day readers (2002:15)[24]. Tissari (2010a:127; 138), in researching metaphors surrounding love in EModE and PDE, finds that although some are the same, changes to the surrounding socio-cultural context such as different dominant moral frameworks cause them to have different potential interpretations. She argues that there has been a shift from a "duty-based code of behaviour" to a "right-based code of behaviour" between then and now (2010a:138). This would affect a contemporary interpretation of the meaning of a metaphor.

From the research mentioned above, it was clear that a systematic and detailed study of metaphor in my results (which would be necessary for a reliable interpretation) would be outside the scope of this study. Nevertheless, it is possible to make some specific links with existing research into metaphor in EModE in the course of my analyses, particularly in 7.4. For more on metaphor in Shakespeare's plays, see also Barcelona Sánchez (1995), Freeman (1995), Oncins-Martínez (2011) and Tissari (2009). Tissari (e.g. 2010a and 2010b) examines metaphors in English from EModE to the present day, and Oncins-Martínez (2006) examines sexual metaphors in EModE using data from dictionaries of the 16th to 18th centuries[25].

In this section (3.3), I have explained that carrying out an automated analysis of semantic domains in EModE plays is not without problems. It is still a worthwhile contribution to my study, however, because:

(i) it tests out the USAS tool with a new and different corpus of EModE plays;

(ii) the results go beyond the lexical and lexico-grammatical level of the results from individual words and word clusters, respectively; and

[24] 21,263 words are given in Crystal and Crystal's 2002 book; the full database of 47,365 words is given on their website, see www.shakespeareswords.com (last accessed 10.08.12).
[25] See also Allan (e.g. 2010) for discussions of metonymy in historical English.

(iii) the results give much more information than could be achieved by manual analysis in providing indicators of similarities and differences in the concepts favoured by Shakespeare and other contemporaneous dramatists in their plays.

In the next section, I explain the practical considerations and decisions taken over the settings and parameters used with the corpus linguistic software tools, in order to optimise the reliability and usefulness of the empirical results of all types in the study.

3.4 Practical considerations in obtaining reliable and useful results with corpus linguistic software tools

In his (2009) study of Shakespeare's *Romeo and Juliet*, Culpeper makes clear that the actual quality of key results and their usefulness to the process of stylistic analysis depend on the researcher fully understanding the implications of various alterable parameters in the corpus linguistic software tools, and the influence of different kinds of reference corpora. He suggests some practical ways for improving the quality and reliability of key results, which also have bearing on the other methods applied in my study. I explain the consideration given to these matters below, and the settings I use (reference corpora are discussed in 3.6).

I begin with the type of statistical significance test used (in 3.4.1), followed by the p value (3.4.2) and minimum/maximum frequency settings (3.4.3). These constrain the number of results generated (given that the possibilities for analysis are limited in any study). As Baker (2004:351-352) points out, however, there are no set rules about the cut-off points and parameters which should be used, since these vary according to the size and contents of corpora and the aims which individual scholars wish to fulfil using their data. From the point of view of linguistic importance, it is hard to argue that some results are significant and yet others are not, just because some occur above a chosen threshold, particularly if those just below it occur with a closely similar p

value or log-likelihood value to those just above. For that reason, absolute keyness values are generally not used by corpus stylistic researchers as a definitive guide to the results which justify further analysis, but as a way of ordering the words and other language features which are most potentially important, on an empirical basis. The researcher can then use the relative statistical values (not the absolute values) as an indicator of which ones might reward closer investigation.

3.4.1 Tests for statistical significance

As indicated in 2.5.2, I use the log-likelihood statistical test (Dunning 1993) for calculating key and locked results in *WordSmith* and *Wmatrix*, since it is the only option in *Wmatrix*. The chi-square test (see e.g. Oakes 1998:24-29) is also an option in *WordSmith*, though Culpeper (2002, 2009) finds that comparisons with both tests produce very similar results. Rayson et al. (2004b:928) argue that the log-likelihood statistical test is more reliable than the chi-square test for expected frequencies below 5 (although such low frequencies are not an issue in my study). I noted some criticisms of statistical significance tests as a measure of keyness in 2.5.3.

3.4.2 Minimum and maximum frequency settings for key and locked results

As indicated in 3.2.1, non-localised results which are potential style markers best serve my aims of comparing Shakespeare's language style to that of other contemporaneous dramatists in the parallel corpus. Culpeper (2009:35-36) discusses the imposition of a minimum frequency cut-off to reduce the number of localised or topical results, i.e. those which are less likely to be stylistically important (see further 2.7). In my study, the minimum frequency needs to be sufficiently high to eliminate, or at least minimise, results which are localised to one or a few plays. These are likely to be topical in the

corpus of Shakespeare's plays, or associated with an individual authorial style in the parallel corpus.

Furthermore, as explained in 2.6, the locked results are by definition those of high frequency. In identifying lockwords in British English, Baker (2011:70) uses a minimum frequency of 1,000, taken from his four corpora overall, each of which is a million words in size. Baker points out, though, that his corpora are still relatively small samples, and therefore that smaller numbers of results from them might not reliably represent language trends in whole language varieties. My corpora are each about 800,000 words (see 4.4), but although they are smaller than Baker's corpora, they are larger samples of what is available of the text-type under investigation. The *SDC* contains Shakespeare's First Folio in its entirety, and an estimate of the *NDC* would be that it contains slightly less than one third of the English drama produced at around the same time (based on Craig and Kinney's 2009:xvii corpus of EModE drama being 3.25 million words; see further 4.3.1).

My tests showed that a minimum frequency of 200 produced a manageable number of results for keywords and lockwords, and also for key and locked semantic domains. For the 3-word cluster results, however, I used a lower minimum frequency of 50, in order to generate sufficient results, since recurrent word combinations occur much less frequently than single words (as pointed out by Mahlberg 2007:12). In principle it is important to use the same minimum frequency for key and locked results of the same type, in order to be able to claim them as statistical opposites (part of Baker's 2011 definition of lockwords; see 2.6). For all the locked results, I raised the maximum frequency parameter in *WordSmith* to its highest, which is 16,000 in version 3.0 (Scott 1999) (as advised by Baker, personal communication, 27.10.11). This is

because the default maximum of 500 risks excluding some of the high-frequency results which qualify as locked. There is no maximum frequency setting in *Wmatrix*.

It is worth noting that the minimum frequency settings in *WordSmith* and *Wmatrix* relate to observed (actual) frequencies. With small numbers of results, the distinction between observed and expected frequencies can affect reliability in corpus studies, as is made clear by Rayson et al. (2004b). The minimum frequencies used in my study are above the levels at which reliability is likely to be compromised, however, so I do not discuss observed and expected frequencies further.

3.4.3 P values for keyness and locking

The log-likelihood cut-off points associated with different p values discussed in this section are taken from Rayson et al. (2004b:7)[26]. In 2.5.3 I noted that p values as a measure for keyness are widely used, though not without criticism, and I argued that substantial existing research indicates that this method nevertheless produces results which are potentially useful for stylistic analysis. I do not discuss the pros and cons of using p values further, other than to explain the choice of the setting I used.

Baker outlines the statistical comparison performed by *WordSmith* in generating a keywords list, and explains the role of the p value as follows[27]:

> The p value (a number between 0 and 1) indicates the amount of confidence that we have that a word is key due to chance alone – the smaller the p value, the more likely that the word's strong presence in one of the sub-corpora isn't due to chance but a result of the author's (conscious or subconscious) choice to use that word repeatedly. (2006:125)

The setting of the p value in keyness studies determines the threshold above which words or other linguistic units will be considered to be statistically significant in the two corpora (in other words, what will qualify as "key" and "locked"). The nearer to 0,

[26] See also http://ucrel.lancs.ac.uk/llwizard.html (last accessed 10.08.12).
[27] *Wmatrix* works in a similar way, as detailed for example by B. Walker (2010:369-370).

the more evidence there is for a difference in frequency between the two corpora (i.e. the more key the result is). The opposite nature of locking to keyness, explained in 2.6, means that the nearer to 1.0 the p value (or the log-likelihood to 0), the less evidence there is of a difference in frequency, and the more locked a result will be.

For key results, other corpus stylistic researchers have used p values of between 0.05 and 0.000001. A p value of 0.05 equates to a log-likelihood cut-off value of 3.84, and a 95% probability that results are key due to chance alone, which is a generally acceptable level in social science studies (Baker 2006:125). A p value of 0.000001 is the lowest possible setting in *WordSmith*, equating to a one in a million probability that results are key due to chance[28]. The decision largely depends on the size of the corpora and the type of language unit being analysed. For example, to obtain keywords for their respective studies of Shakespearean dialogue, Culpeper (2002) uses a p value of 0.05 with a corpus of c. 20,000 words from one play, whereas Scott and Tribble (2006) use a p value of 0.000001 with a corpus of 37 Shakespeare plays amounting to c. 800,000 words. Corpus stylisticians working with prose fiction have found other p values to be satisfactory: B. Walker (2010:370) opts for a p value of 0.001 to obtain keywords, key parts of speech and key semantic domains from his 73,000-word corpus of a novel, and Mahlberg (2007:12) uses a p value of 0.00001 to extract key 5-word clusters from her 4.5 million-word Dickens corpus.

There are no precedents for determining the principle of a cut-off point for locked results, i.e. how far below a p value of 1.0 can a result still be considered to be locked. *WordSmith's* output shows a keyness value for any item which has a log-likelihood value of 1.0 or above, but for items with a lower log-likelihood value only the p value is shown. For the purposes of this study, I therefore take a log-likelihood

[28] See also Scott (e.g. 1999:Help menu).

value of 1.0 as the point below which results become locked. This equates to a p value of about 0.3. There may well be cases for applying other principles, however (for example, that an equivalent set of distances from p=1.0 should be established for lockwords as for keywords, such as those in the studies mentioned above). However, the testing and evaluation of these ideas is well beyond the scope of this study, and is not necessary to achieve my research aims. The principle I follow enables me to generate ordered lists of the most locked items, from which I can investigate stylistic similarities between the two corpora. The statistical basis of "locking" would benefit from further exploration and discussion in future research, however.

I found that even the lowest possible p value of 0.000001 in *WordSmith* (which corresponds to a log-likelihood value of approximately 27 in *Wmatrix*) produced more than sufficient keyword and key semantic domain results to discuss in the space available. Since the number of key word cluster results generated were fewer, as mentioned in 3.4.2, for these results I used a p value of 0.01 (log likelihood=6.63) as the cut-off threshold. A p value of 1.0 (log-likelihood=0) generated more than adequate numbers of locked results for words, word clusters and semantic domains. The cut-off thresholds for p value and minimum frequency effectively help to manage the distribution of results, an aspect of corpus output which I discuss in more detail in the next section.

3.5 Distribution of quantitative results

When analysing quantitative results, it is important to assess their distribution across the contents of the corpora (Baker 2004:350). This is to see whether the source of their quantitative significance is located in particular areas of the corpus (for example, in certain plays, in particular genres, or near the beginning or the end of plays). An

awareness of the way results are distributed throughout the corpora is therefore crucial in order to avoid overstating or over-generalising the extent of their effects in EModE plays. Style features which occur as quantitatively significant may be used relatively frequently by only **some** of the characters whose dialogue is represented by the results in a particular dataset. For example, a result which is positively key in Shakespeare's plays may turn out to be concentrated more heavily in the comedy genre only, or in the dialogue of characters of one gender.

I checked the range of plays in which my results occur using dispersion plots from *WordSmith* for word and word cluster results, and from *AntConc* for semantic domains (by exporting the *Wmatrix* concordance data, as explained in 3.3.2 above). As discussed in chapter 2, I focus on high-frequency items in the corpora. The majority of the results proved to be distributed fairly evenly across the plays, and across the three genre components. Therefore, in order to avoid excessive repetition, in my analyses in chapters 6 to 8 I comment mainly on distribution patterns which are **uneven**. I take distribution as one factor which guides my decisions about results which are of most interest to follow up with closer analysis, and I include distribution plots for particularly interesting results (but not routinely).

It is worth noting that the distribution of results can be quantified precisely, if required, using methods such as those applied by Leech et al. (2001:18-19) to the *BNC*. They use the following "dispersion indices":

- "Range": the "number of sectors of the corpus (out of a maximum of 100) in which the word occurs".
- "Dispersion": "a statistical coefficient (Juilland's D) of how evenly distributed a word is across successive million-word sectors of the corpus" (2001:18).

These methods would be quite time consuming to apply, however, and the level of detail greater than is necessary to assess the distribution of results across my much smaller corpora comprising only one text-type (EModE plays).

3.6 Reference corpora

The extent to which the reference corpus influences the results in corpus studies is the subject of debate. In this section, I briefly discuss the main issues in corpus linguistics regarding sizes and contents of reference corpora, and make clear how these affect my decisions in building the *NDC* as a specialised reference corpus for Shakespeare's First Folio (which are detailed in the next chapter). In the course of my earlier discussions of keyness studies, in 2.3 and 2.5.2, I have mentioned some reference corpora used in other corpus stylistic research. Although existing research is oriented to keyness studies, the content and choice of the reference corpus has implications for the opposite statistical concept of locking. Scott (2010:43) points out that keyness, as a corpus linguistic concept, is "text-dependent" rather than "language-dependent". This is also true of locking. The results that are generated from texts in a corpus analysis depend to some extent upon what they are being compared with.

With regard to the content of reference corpora, Culpeper (2009:35) makes the important observation 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". It is important to choose a reference corpus which is appropriately close or distant in content to the one under investigation, depending on the kind of claims that are to be made from the results. A good example of this is provided by Fischer-Starcke (2009), who uses three different reference corpora to obtain results which illuminate different aspects of language style in Jane Austen's

novels. Fischer-Starcke compares the single work *Pride and Prejudice* with a reference corpus of Austen's other five novels, to identify language features which are particular to that novel in the context of the writer's other works. She gains a broader perspective on novel-specific language features by comparing *Pride and Prejudice* to a reference corpus of other contemporaneous fiction, and finally she investigates authorial style features in Austen's works by comparing the corpus of all six Austen novels to the corpus of other contemporaneous fiction. Fischer-Starcke notes that some results occur as key with more than one reference corpus, which she analyses in closer detail as the strongest evidence for Austen's style (2009:498-499). Her choice of narrow and then broad reference corpora enable more certainty over results that reflect Austen's individual style, rather than general trends among novelists of the period.

In my study, I aim to make fine distinctions between texts of the same type (EModE plays) and from the same dramatic genres; the variable which is in focus is the author's style (Shakespeare's compared to that of a group of his peers). In order to show how Shakespeare's language style is similar or different to that of his peers, the content of the reference corpus must be closely similar (i.e. parallel) to that in Shakespeare's First Folio, otherwise the results (keywords in particular) may reflect other factors. To illustrate the point, comparing Shakespeare's First Folio to a reference corpus of PDE such as the *BNC* would be likely to produce key and locked results reflecting aspects of language which have or have not changed in English over 400 years. Comparing it with a reference corpus of speech-related text from the Early Modern period (such as the *CED*) would be likely to produce key and locked results which show differences and similarities between drama and other registers or text-types from the same period. It is still possible that these comparisons would yield statistical results which reflect Shakespeare's style features, but the point is that there

would be room for doubt. This can be much reduced if the reference corpus is narrowed to one of closely similar content to Shakespeare's First Folio, without actually containing Shakespeare's plays.

With regard to size, while some scholars claim that a reference corpus should be much larger than the corpus under investigation (e.g. Berber Sardinha 1999), Scott's (2009) findings suggest that the content of the reference corpus is more crucial than the size, where only one register or text-type is being investigated. Scholars working with literary texts have used different approaches, which suggest that different kinds of reference corpora are suitable for particular research aims. Culpeper (2002, 2009) uses all the other characters' dialogue as a reference corpus for the dialogue of each character he investigates in *Romeo and Juliet* (see 2.5.2). In contrast, Scott and Tribble (2006) use all Shakespeare's plays as a reference corpus in their study of *Romeo and Juliet*, without excluding the text of the play or character under consideration. They note that Culpeper's approach leads to results which "home in on individual difference", whereas theirs tends to yield a "robust core" of results (Scott and Tribble 2006:64).

There are further precedents for using reference corpora of either equal or much larger size in corpus studies of literary works by one author. Mahlberg's (2007) corpus of (non-Dickensian) 19th century fiction is of similar size to her Dickens corpus (about 4.5 million words). In contrast, Fischer-Starcke's (2009:497) corpus of literature contemporaneous to Jane Austen's work, which is also over 4 million words in size, is much bigger than the 602,000-word Austen corpus. As also mentioned in 2.3, Leech (2008) uses fiction by multiple contemporaneous authors as a reference corpus for Woolf's *The Mark on the Wall*, though he emphasises the comparability of content (genre and historical period) rather than the relative sizes of the corpora.

Though there is no definitive ideal, the above studies make clear that the content of a reference corpus has bearing on the nature of results extracted from the corpus under consideration, and on the kinds of claims which can reliably be made based on those results. There are also practical decisions to be made about the size, in view of the availability of texts which are suitable for comparison. As I discuss in 4.3.1, for his analyses of Shakespeare's language style, Culpeper (2011) uses a very large reference corpus of EModE drama which includes Shakespeare's plays. In my study, I use a much smaller but more specialised corpus which parallels Shakespeare's First Folio in size, date, genre and other content features as closely as is feasible, because I want to maximise the possibilities of obtaining data reflecting Shakespeare-specific language style. It is interesting to note similarities between Culpeper's (2011) results and mine (in 6.3). Nevertheless, though some of the same results might be obtained using other, broader reference corpora, the parallel reference corpus constructed for this study affords more certainty that the results reflect Shakespeare's authorial style in relation to that of other contemporaneous dramatists. It therefore enables me to address research question 1, concerning language styles in plays by Shakespeare and other contemporaneous dramatists, in the most reliable way. A larger reference corpus would inevitably contain EModE plays that are more distant from Shakespeare's in date and other criteria. As I show in the next chapter, it is actually quite challenging to construct a parallel corpus of equal size to the First Folio.

3.7 Summary

In this chapter, I have provided some further detail about the linguistic items that, in addition to single words, constitute the quantitative results from which my qualitative analysis proceeds (word clusters and semantic domains, in 3.2 and 3.3 respectively). I

have argued that they are relatively under-researched, and that they add value to my investigation of language styles in Shakespeare's plays and other contemporaneous plays, based on existing studies such as Archer et al. (2009) and Culpeper (2011) and the arguments of Crystal (2008). My investigations also test these methods further with EModE, since they are still quite new. I have explained the anticipation and handling of some problems and limitations with their use.

I have also set out my rationale for the settings of the corpus linguistic software tools, which are effectively the operational definitions of "key" and "locked" results (in 3.4). I have argued that these produce the most reliable and useful results for my analysis, whilst keeping the numbers at a manageable level. I have mentioned the importance of analysing distribution in order not to over- or under-state the implications of results, and I have explained how I present this in my analyses (in 3.5). In 3.6 I introduced briefly the issues surrounding reference corpora, and argued that closeness of content is crucial to claiming that results can be said to be indicative of language styles in the corpus under consideration (as opposed to the general register or text-type from which it comes). This underlies my decision to construct a parallel reference corpus for Shakespeare's First Folio, to maximise the potential for finding evidence of style features which are particular to Shakespeare (and not to EModE plays in general). I discuss this in detail in the next chapter.

My discussions in this and the previous chapter show that corpus linguistic methods offer huge advantages in tracing patterns of linguistic constructs in large bodies of text such as collections of plays, and also that there are also some limitations in the reliability and usefulness of the results. The output is a systematically-obtained guide to what might reward further analysis, but any statistically-generated language profile represents one of many possible angles from which a text or set of texts may be

illuminated. In my study, I have chosen a constellation of methods (simple frequency, keyness and locking, applied to single words, word clusters and semantic domains) to view Shakespeare's language style in the context of that of his peers from a variety of angles, in order to obtain a more well-rounded picture than would be obtained from just one type of analysis.

CHAPTER 4. BUILDING HISTORICAL CORPORA: SELECTING SUITABLE COLLECTIONS OF PLAYS BY SHAKESPEARE AND OTHER CONTEMPORANEOUS DRAMATISTS

4.1 Introduction

This chapter addresses the first two parts of research question 3, concerning the issues that arise in building a historical corpus. I discuss and explain the choices made and the problems encountered in choosing and compiling corpora of EModE plays that enable me to meet the research aims set out in 1.2. As discussed in 1.1, 2.3 and 3.6, my study requires a corpus of Shakespeare's plays (the *SDC*) and a specialised parallel corpus containing a similar amount of dialogue from plays of closely comparable types and dates by other EModE dramatists (the *NDC*).

The background and content of the existing Shakespeare corpus used for the *SDC* is discussed in 4.2. The compilation of the *NDC* is detailed in 4.3. Here, I explain the rationale and inclusion criteria for selecting the most suitable play-texts to compare with those by Shakespeare, from the sources available and bearing in mind some compatibility issues with the texts in the *SDC*. The *NDC* was compiled in late 2009 and comprises a multi-authored collection of plays, chosen from a range of possible options, within various constraints and considerations. I document these in some detail, to demonstrate my efforts to achieve a balanced and representative corpus of plays which is comparable to Shakespeare's First Folio. By this I mean:

- balanced in the proportion of dialogue from each of the three main dramatic genres and from earlier and later periods in which Shakespeare was writing;
- representative of a range of the most popular and successful plays that are contemporaneous with those of Shakespeare; and
- comparable in terms of other factors such as settings (classical and pastoral, for example) and intended audiences (public theatres, plus court performances).

This is to maximise the chances of obtaining results which reflect authorial style features, rather than language features which are only associated with the style of a particular genre or trends in earlier and later dates (since language styles in plays are already known to vary according to those factors). For more on balance, representation and comparability of texts in corpora, see for instance Biber (1993), Leech et al. (2009:28) and McEnery et al. (2006:59-60, 125-130).

Following my discussions of the plays selected for the *NDC*, in 4.4 I provide a table showing a side-by-side comparison of word counts for the main components of the *SDC* and the *NDC*, with some commentary on the extent to which the content of the *NDC* successfully parallels that of the *SDC*.

4.2 Issues surrounding the corpus of Shakespeare's plays

I begin this section by explaining the limitation of the study to Shakespeare's First Folio, in 4.2.1, followed by the source and background of the *SDC* play-texts, in 4.2.2. I list the plays comprising the *SDC* in 4.2.3, where I also introduce some problems surrounding dating and genre.

4.2.1 Plays included in the *Shakespearean Drama Corpus*

In a study which concerns authorial style distinctions between Shakespeare and a group of his peers, it is desirable to minimise the number of plays which are known or suspected to be collaborations between Shakespeare and other dramatists. This is to ensure that the dialogue is as distinctively "Shakespearean" or "non-Shakespearean" as possible. Therefore, I include only the 36 works in the First Folio (of 1623) in the *SDC*. Other plays which appear in subsequent folios and editions include:

- *Pericles* and *The Two Noble Kinsmen* (in modern Oxford editions of Shakespeare's plays such as Greenblatt et al. 1997);
- *Edward III* (in the New Cambridge Edition and the Oxford Shakespeare second edition, as noted by Watt 2009:116); and
- *Double Falsehood* (recently published by Arden Shakespeare; see Hammond 2010).

Having limited the *SDC* to the plays in the First Folio, it is nevertheless still likely that some of the dialogue is the result of collaboration between Shakespeare and other dramatists (in, e.g., *Henry VIII*, *Timon of Athens* and *Titus Andronicus*). Crystal and Crystal point out that collaboration was not unusual among dramatists at the time (2005:57), and they argue that ten or more of the plays may be collaborative (2005:174). As records of authors and contributors to drama were not always made at the time of construction (Vickers 2002:10), exactly how much dialogue was penned by Shakespeare himself remains uncertain. Further discussion of the authorship of EModE drama is given in, e.g., Craig and Kinney's (2009) edited volume, Hope (1994), Petersen (2010) and Vickers (2002). At the present time, however, Shakespeare is still considered to be the sole or main author of plays in the First Folio.

4.2.2 Background to the *Shakespearean Drama Corpus* play-texts and assessment of their suitability for the study

The corpus of Shakespeare's plays on which the *SDC* is based was built by Mike Scott and is publicly available[29]. Scott built his Shakespeare corpus from the 1916 edition of

[29] See http://www.lexically.net/wordsmith/support/shakespeare.html (last accessed 10.08.12).

The Oxford Shakespeare, edited by W.J. Craig, which is also publicly available from the *Online Library of Liberty*[30]. Scott's corpus includes:

- a text file for each of the 36 plays in the First Folio, plus *Pericles*;
- a text file of each individual speaking character's dialogue for every play; and
- a list of dramatis personae for each play.

Scott's Shakespeare corpus is already annotated to make the play-texts suitable for investigation using corpus linguistic software tools. Non-dialogic text such as stage directions is isolated within angle brackets in order to exclude it from computations, whilst still allowing it to be viewed later on to assist in the interpretation of results. Character identity tags are preserved in the text files of whole plays (although not in the individual character files), so as to identify who is speaking. This is illustrated in example (1) (the percentage figure indicates how much of the play has taken place).

(1)
```
<STAGE DIR>
<Enter Oliver.>
</STAGE DIR>
<OLIVER>    <2%>
</OLIVER>
      Now, sir! what make you here?
```

Shakespeare, *As You Like It,* I:i (Scott's Shakespeare corpus)

The ability to access the kind of socio-pragmatic information which is tagged in Scott's corpus is important, because dramatic dialogue (at least, that in Shakespeare's plays) is mainly interactional[31]. Therefore, the identity of the addressee and other co-present individuals has bearing on what speaking characters say. This is argued by Culpeper (2001:167-172), who finds that one way that dramatists characterise individuals in plays is through "self-presentation" and "other-presentation", i.e. what

---

[30] A collection of literary works available free online for academic and personal use. See http://oll.libertyfund.org (last accessed 10.08.12).
[31] For more on the interactional nature of drama, including Shakespeare's plays, see Herman (1995).

the characters say about themselves and others, to (or in the presence of) other people in the play.

During the course of my (2009) corpus research, I adapted Scott's corpus by adding further annotation to the text files of each character's dialogue (over a thousand in total; see Demmen 2009:49-66). I added tags for speaker names and the titles of the plays, so that I could use these files to build corpora of male-only and female-only dialogue, and pinpoint exactly which character was speaking when I came to examine my results. Example (2) shows the text in example (1) above in the form it appears in my adapted version of Scott's corpus (the first two tags are the ones I added).

(2)
```
<OLIVER><AYLI><SPEECH 1><ACT 1><SCENE 1><2%>
Now, sir! what make you here?
</OLIVER>
```

Shakespeare, *As You Like It,* I:i (*SDC*)

In addition to further annotation, I also regularised the EModE spelling variation in the corpus using Baron and Rayson's (2008) *VARD 2* software, thereby increasing the potential for retrieval of results derived through automated orthographic matching processes (such as frequency lists and keyness, explained in 2.5; I discuss *VARD 2* further in 5.4). The specialised annotation and regularised spelling meant that my version of Scott's Shakespeare corpus required little further treatment to work with the methods used in the present study (explained in 2.5 and 2.6). The text of *Pericles* was removed, so it contains only the First Folio, for reasons stated in the previous section, and untagged files were created for use with *Wmatrix* (see 3.3.2). Using this corpus advantageously maximised the time for building the new *NDC* and preparing the texts for investigation with corpus tools.

Apart from the slight inconvenience of the non-XML tags being incompatible with *Wmatrix*, the only disadvantage in using Scott's (adapted) corpus in this study is that it is based on a 1916 edition of Shakespeare's plays, whereas the *NDC* texts are

from sources dated about 300 years earlier (obtained from *EEBO*, as explained in 4.3). The *SDC* therefore contains an edition which has undergone some modernisation of the language, whereas the *NDC* contains early extant editions which have not. There was no way of getting around this in the study, because there were insufficient modernised versions of other contemporaneous plays from which to construct a parallel reference corpus for Shakespeare's plays, and not enough time in the project to construct and/or prepare a corpus of early extant Shakespearean play-texts in addition to the *NDC*. This causes some minor compatibility problems between the language of the two sets of texts, which in a few cases affects the results, particularly those which contain punctuation (e.g. in 8.2). I assessed the likely impact of this at the start of the study by comparing samples of text from my adapted version of Scott's corpus with corresponding text samples from a digitised early extant edition of the First Folio (dated 1623 and downloaded from *EEBO*). Variation between the 1623 and 1916 editions was mainly evident in spelling and punctuation, as illustrated by examples (3) and (4) (with variation highlighted in bold type).

(3)
(1623)

An.
Friends, Romans, **C**ountrymen, lend me your ears:
I come to bury **C**aesar, not to praise him:
The **euill** that men do, **liues** after them,
The good is oft **enterred** with their bones,

(4)
(1916)

<ANTONIUS> <59%>
Friends, Romans, **c**ountrymen, lend me your ears;
I come to bury **C**æsar, not to praise him.
The **evil** that men do **lives** after them,
The good is oft **interred** with their bones;

Shakespeare, *Julius Caesar*, III:ii

Other samples occasionally showed grammatical variation and variation in the lineation. Historical spelling and punctuation are not the focus of my study, and spelling is regularised in the corpora in any case, in order to improve the potential for

retrieving results with the corpus tools (as mentioned above). Moreover, as discussed in further detail in 5.4 and 5.5, spelling and grammar practices vary even between early extant texts of similar date and source in this historical period (such as the use of apostrophes in contractions). There are two reasons for this:

(i) English spelling was not fully standardised until around 1650 (according to Nevalainen 2006:32), i.e. some 20 to 50 years later than the construction period of the texts in this study; and

(ii) Variation in spelling and punctuation reflects the preferences of different compositors of the printed editions of plays, which are not necessarily the preferences of the dramatists.

The above reasons mean that it would be difficult to make reliable claims about language styles associated with spelling or punctuation features in any case. I concluded that a study of grammar would be more affected by a comparison of early and later editions such as those in the corpora used in this study (through editorial changes to word endings, for example). The kinds of grammatical aspects relevant to my investigations are at word, phrase or sentence level (for example, choices of pronouns and formulaic constructions such as utterance launchers, mentioned in 3.2.1). My assessment of the 1623 and 1916 editions of Shakespeare's plays indicated that these remain substantially the same.

The question of how much modernisation is acceptable in texts used for historical linguistic studies is not easy to answer, and depends on the aims of the researcher and the methods used. Culpeper and Kytö argue that modernised editions might be based on an amalgamation of earlier printed editions of a play, so they adopt the principle that "the date of printing should be close in time to the date of creation of the text" in constructing the *CED* from the earliest extant editions (2010:26). This is

critical in their historical sociolinguistic study, because the dialogic data needs to be as close as possible to naturally-occurring speech in the Early Modern period. My aims are different, however, being more focused on the stylistics of drama rather than on historical aspects of speech-related language (though there are connections, of course, as noted in 1.2).

Aside from Scott's Shakespeare corpus, other digitised collections of Shakespeare's plays exist, which have been used by scholars according to the needs of their research. For example, Murphy (2007) uses *The Nameless Shakespeare* (Mueller 2005) which is accessed through the computer interface *WordHoard* (Mueller et al. 2006). Although this is a powerful programme for investigating morphological, syntactical, semantic and narrative information in the plays, it does not easily provide the kind of socio-pragmatic information I require in this study. B. Busse (2006) and U. Busse (2002a) use Spevack's (1968-1980) concordances (to which I did not have access), for their corpus studies of vocatives and pronouns, respectively, in Shakespeare's plays. Culpeper (2002:14) sources his text of *Romeo and Juliet* from Craig's 1914 edition of *The Oxford Shakespeare*, which is publicly available from Bartleby's online collection of literary and classical works[32]. Different user interfaces and annotation systems suit different kinds of corpus tools and methods, because there are compatibility issues among them (as discussed further in 5.2).

4.2.3 List of plays in the *Shakespearean Drama Corpus*

The plays in the *SDC* are listed in chronological order in Table 3 (comedies), Table 4 (histories) and Table 5 (tragedies). Following the tables, I highlight some issues regarding dating and dramatic genres, which are discussed further in 4.3. Word counts

---

[32] See http://www.bartleby.com/70/ (last accessed 03.09.12).

for each of the plays are also given in the tables (from *WordSmith*), and the total number of words of dialogue in the *SDC* is 797,054.

Table 3. Comedy plays in the *SDC*

Play title	**Date of first construction, performance, printing or publication**	**Date of edition in corpus**	**Total dialogue word count**
COMEDY PLAYS			
The Comedy of Errors	1592-1594	1916	14,630
The Taming of the Shrew	1593-1594	1916	20,519
Two Gentlemen of Verona	1594	1916	16,960
Love's Labour's Lost	1594-1595	1916	21,048
A Midsummer Night's Dream	1595-1596	1916	16,006
The Merchant of Venice	1596-1597	1916	20,982
The Merry Wives of Windsor	1597	1916	21,342
Much Ado About Nothing	1598-1599	1916	20,883
As You Like It	1599	1916	21,330
Troilus and Cressida	1601-1602	1916	25,418
Twelfth Night	1601-1602	1916	19,543
All's Well That Ends Well	1602-1603	1916	22,661
Measure for Measure	1604	1916	21,376
Cymbeline	1609-1610	1916	26,976
The Winter's Tale	1610-1611	1916	24,786
The Tempest	1611	1916	16,156
Total Shakespearean comedy dialogue			330,616

Table 4. History plays in the *SDC*

Play title	**Date of first construction, performance, printing or publication**	**Date of edition in corpus**	**Total dialogue word count**
HISTORY PLAYS			
Henry the Sixth Part One	1589-1590	1916	20,628
Henry the Sixth Part Two	1590-1591	1916	20,882
Henry the Sixth Part Three	1590-1591	1916	23,389
Richard the Third	1592-1593	1916	28,391
King John	1594-1596	1916	20,484
Richard the Second	1595	1916	21,884
Henry the Fourth Part One	1596-1597	1916	24,080
Henry the Fourth Part Two	1598	1916	27,168
Henry the Fifth	1599	1916	25,704
Henry the Eighth	1612-1613	1916	23,161
Total Shakespearean history dialogue			235,771

Table 5. Tragedy plays in the *SDC*

Play title	**Date of first construction, performance, printing or publication**	**Date of edition in corpus**	**Total dialogue word count**
TRAGEDY PLAYS			
Titus Andronicus	1593-1594	1916	21,413
Romeo and Juliet	1595-1596	1916	23,855
Julius Caesar	1599	1916	19,173
Hamlet	1600-1601	1916	29,832
Othello	1604	1916	26,039
King Lear	1605	1916	25,396
Macbeth	1606	1916	16,575
Antony and Cleopatra	1606-1607	1916	23,874
Coriolanus	1607-1608	1916	26,698
Timon of Athens	1607-1608	1916	17,812
Total Shakespearean tragedy dialogue			230,667

There is debate over the genre into which some of Shakespeare's plays fit (see for example Boyce 1990:119-120, 652-653 and Crystal and Crystal 2005:33, 83; see further Hope 2010:170-205). I use the classifications of Greenblatt et al. (1997), which are conventional in modern editions. *Troilus and Cressida* and *Cymbeline* are classified as comedies, though as Boyce (1990:653) notes, they are listed as tragedies in the First Folio. I do not formally sub-categorise "early", "middle" and "late" comedies, "romances" and "problem" or "tragicomedy" plays (further details of these distinctions are given in, e.g., Boyce 1990:119-120).

The dates of construction in the tables above are from U. Busse (2002a:43-44), who acknowledges some differences of opinion about them. Busse uses scholarly sources for dating (mainly Wells et al. 1987:69-44, with supporting information from others such as Evans and Tobin 1997:78). Dating of plays, and indeed other texts from the Early Modern period, is a thorny issue which creates difficulties for researchers. Gaps in historical records mean that dating often cannot be verified with certainty, and there are debates over the quality of evidence for dating among critical editions of plays. The early extant editions are variously based on the date of construction of a

play, its first performance, first printing or first publication. This is further complicated by time lapses between these events in some cases, particularly publication, and by the absence of complete records.

The lack of dating information, combined with the complication of collaborative playwriting practices (noted in 4.2.1 above), reflect the fact that the whole concept of "authorship" was not nearly as fixed in the Early Modern period as it is in the present day. This is discussed by Petersen (2010:3-34), who argues that "key primary sources like Henslowe's *Diary* and the Stationers' Register [...] chart a history of collective practice, anonymous publication, random attribution, and an at best irregular (nascent) concept of 'copyright'" (2010:17). The Stationers' Register was a record of written works held by the Stationers' Company (the guild which regulated publishing and printing trades) under a royal charter (see further Dutton 2000, e.g. 99, 104 and Ioppolo 2002:171-173). Writers and publishers often entered their works in the Stationers' Register to assert their rights over them, but this did not necessarily happen in the same year as the works were first constructed, first performed or printed.

With regard to the issue of "good" and "bad" editions of EModE plays, Crystal (2008:22-23) argues that "[a]ll texts of the period, regardless of their literary status, must reflect some sort of contemporary linguistic practice, so everything is of value." I tend to agree with this, particularly in view of limited sources of Early Modern language data, although it is important to evaluate the sources and note any issues which might affect the study, as I do in the next chapter with regard to the texts for the *NDC* sourced from *EEBO*. Although an inconvenient compromise for linguists working with EModE drama, it is a necessary one, because literary and linguistic research (Greenblatt 1997:67 and Petersen 2010:13, respectively) indicates that the notion of one definitive "original" version of a play-text is actually a fallacy. This is

because successful plays (which are of most interest for study) were rapidly changed and amended by dramatists and acting companies to suit different performance locations and types of audiences. Petersen states that:

> As a cultural phenomenon, the early modern stage was an almost unique event in the history of popular entertainment in England. [...] In this transitionary space between genuinely traditional live entertainment and literary drama, there is not yet a stable concept of what might call an *accurate* text, but rather there are many versions and many kinds of text in circulation, filling *adequate* textual roles (artistic and practical; functional, social and political) in the traditions of individual plays. (2010:7, Petersen's emphasis)

The abovementioned issues mean that some dates of plays given in this study are approximate, and based on the most reliable sources I could access. For all the play-texts in the study, I use the year of construction where possible, but where this is unverifiable I use the year of first performance or year of first printing, in that order of preference, depending on the information available. I give the year of publication separately.

In this section (4.2) I have explained the source and contents of the *SDC* and highlighted the slight disadvantage (in this study) of its modernised texts. I have also introduced some of the problems which are inherent to studying historical texts: verifying dates, describing dramatic genres, and authorship attribution. These become even more apparent in my detailed discussions of the construction and preparation of the *NDC*, in the next section.

4.3 Issues surrounding the corpus of other contemporaneous plays

In 4.3.1, I discuss potential sources of EModE play-texts which I investigated for the *NDC*. Then, in 4.3.2, I explain the selection and rejection of play-texts on the basis of balancing dates and genres, together with other factors which make them relatively

more or less comparable with Shakespeare's plays and/or representative of the types of popular drama of the period. A list of plays in the *NDC* is given in 4.3.3.

4.3.1 Sourcing play-texts for the *Non-Shakespearean Early Modern English Drama Corpus*

At the start of the study, I investigated whether some or all of the content of existing corpora containing EModE drama could be used or incorporated into a parallel corpus for Shakespeare's plays, but found that these were either not publicly available and/or not suitable for my chosen methods. An early corpus of EModE plays built by Eleanor Mitchell (1971) for her PhD is not publicly available, though is worth noting as one the first of its kind, constructed without the advantage of downloadable material from the World Wide Web (and see further U. Busse 2002a:49-52). The *CED* (mentioned in 2.3) is accessible at Lancaster University. In its drama section there are 25 comedy samples by playwrights other than Shakespeare (as well as a Shakespearean sample). All the samples are annotated so that non-dialogic text is excluded from computations. However, I rejected this as a source, for two reasons:

(i) the samples are not sufficiently large or diverse to constitute a parallel reference corpus for the whole of Shakespeare's First Folio; to achieve sufficient data requires entire play-texts, from all three genres; and

(ii) some of the samples are dated too early or too late to be comparable to Shakespeare's plays.

The 3.25 million-word corpus of EModE drama used by Craig and Kinney (2009) and their colleagues for computational stylistic research into authorial styles is not publicly available, and their inclusion criteria are also broader than mine. They use all the non-Shakespearean plays that are available, as long as they are dated between 1580 and 1619 and authored by one person only (2009:xvii). As indicated in 4.1 and discussed

further in 4.3.2, to construct a **parallel** corpus for Shakespeare's First Folio it is necessary to consider other factors as well as dates. Hope and Witmore's (2010) corpus is also not publicly searchable, and again their inclusion criteria are much broader than mine. Their corpus contains 320 plays spanning the period 1519-1659 (2010:388), and they use different methods to those in my study (see 3.3.1).

The *Korpus of Early Modern Playtexts in English* ("*KEMPE*"), used in computational stylistic research by Petersen (2010), is publicly searchable. It comprises 287 plays from the Early Modern period, amounting to just under 9 million words (Petersen 2010:164, 278-305). *EEBO* and the online literary archive *Project Gutenberg* are listed among the electronic sources, together with *Chadwyck Healey Literature Online* ("*LION*") (2010:277). However, some of its features make it unsuitable for my study. I could not find a way of selecting some parts of the corpus but not others, so it would not be possible to isolate plays by Shakespeare and exclude them from the rest. This is also found by Culpeper (2011:80, note 9), who uses *KEMPE* as a reference corpus for Shakespeare's plays in his comparison of lexical bundles in Shakespearean drama, other contemporaneous drama, EModE courtroom trial data and PDE drama. That is not necessarily problematic: as discussed in 3.6, Scott and Tribble (2006) include the text under consideration as part of the reference corpus in their study of Shakespeare's plays, and find that a similar set of results is produced regardless of whether or not it is excluded. However, the Shakespearean content of *KEMPE* is not only included, but duplicated to some extent, since it contains 21 quarto editions as well as 36 folio plays by Shakespeare (Petersen 2010:175). Although the folio and quarto editions are probably not identical, some of Shakespeare's plays are essentially doubly represented in *KEMPE*, compared to others. That presents a risk of skewing results based on statistical comparisons such as

frequency, keyness and locking, through exaggerating the effects of language style features associated with particular play(s) or character(s). The amount of exaggeration may not materially alter the distinction between Shakespeare's language style and that of other contemporaneous dramatists in studies such as mine and that of Culpeper (2011), but the point is that there is potential for it to do so, and the reliability of statistically-based results is therefore less certain. In contrast, in a study concerned with authorship attribution, such as Petersen's, more data from any one author usefully provides more evidence on which to base claims. Furthermore, the EModE spelling in *KEMPE* is in its original form (Petersen 2010:164). This would reduce the potential for retrieving results using the methods in my study, which rely on orthographic matching (see further 5.4).

The above assessment of existing corpora led to my decision to create a new reference corpus of other contemporaneous drama to compare with Shakespeare's First Folio in this study. A parallel corpus of equal size and closely-similar content to Shakespeare's First Folio – whilst not actually duplicating any of it – provides a more reliable basis on which to make claims about language style features which are or are not shared by Shakespeare and other contemporaneous playwrights. As noted in 4.2.2, there were insufficient sources of digitised modernised EModE play-texts by dramatists other than Shakespeare from which to compile the *NDC*[33]. In one sense this would have offered an advantage through greater compatibility with the texts in the *SDC* but, in another, it would have been a potential disadvantage. This is because the modernisation of play-texts with many different editors would be inconsistent and unverifiable (unlike that in the play-texts in the *SDC*, whose texts are based on a single

[33] A few are downloadable from *Project Gutenberg*.

edition from one scholarly source, with editing that is standardised)[34]. Copy-typing hard copies of modern editions of EModE plays, or scanning them in with optical character recognition ("OCR") software and checking them, would have been too time-consuming (a limitation also faced by Culpeper and Kytö 2010:26), and hampered by copyright restrictions. Therefore, the *NDC* is compiled from early extant play-texts dating from the late 16th and early 17th centuries, downloaded from *EEBO* (to which Lancaster University subscribes).

A great many plays from the Early Modern period survive[35], notably in the *English Short Title Catalogue (1475-1640)*, compiled by A.W. Pollard and G.R. Redgrave and published in 1927, and the *Short-Title Catalogue 1641-1700*, compiled by Donald Wing and published between 1945 and 1951[36]. These are both accessible through *EEBO*. The texts are available as files of microfilmed printed manuscripts, many of which are now also available as digitised typescript text files. These are being transcribed through the ongoing "Text Creation Partnership" between the University of Michigan, Oxford University and ProQuest LLC, together with other participating libraries and institutions ("*EEBO-TCP*")[37]. I use the earliest edition of each play available on *EEBO* in a digitised format, following Culpeper and Kytö's (2010:26) criterion that dates of printing and construction should be as close as possible (for the drama component of the *CED*, mentioned in 4.2.2). In the next section, I go on to chart the evaluation process of the available digitised play-texts, to select those most suitable for inclusion in the *NDC*.

---

[34] For more on the modernisation of the language in the Oxford Shakespeare, see Wells et al. (1987:155-157).
[35] Petersen (2010:4) cites Gurr's (2000:17) number of 556 extant plays from between 1584-1642, out of two to three thousand that once existed (from details of theatrical history recorded by Philip Henslowe, in Greg 1908:146, vol. 2).
[36] See http://eebo.chadwyck.com/about/about.htm#aboutstc (last accessed 17.07.12).
[37] See http://eebo.chadwyck.com/marketing/about.htm and http://www.textcreationpartnership.org/ (both last accessed 17.07.12).

4.3.2 Identifying suitable contemporaneous plays for comparison with Shakespeare's First Folio

There are a great many digitised EModE plays on *EEBO*, and it would not take long at all to amass about 800,000 words of dialogue by dramatists other than Shakespeare. However, not all of it is sufficiently comparable with the plays in Shakespeare's First Folio to be suitable for a parallel corpus. Date and genre are the most crucial factors in the choice of plays included in the *NDC*, as I discuss in more detail in 4.3.2.1 and 4.3.2.2 below, respectively. As indicated in 4.3.1, other issues surrounding the content or context of play-texts are also considered. Some of these issues arise in the course of my discussion of genres of play-texts in 4.3.2.2, including:

- the sub-genre and topic of the plays, and the types of characters in them;
- dramatists with idiosyncratic language styles;
- plays written for child or adult companies of actors;
- queries over authorship;
- temporal locations and geographical settings of the plays; and
- the numbers of male and female characters in the plays, and proportional quantities of their dialogue.

In 4.3.2.3, I mention a few further considerations.

The principles of careful selection I apply are informed by those given to the construction of the comparative diachronic Brown family of corpora, to ensure parity in the content of each of the corresponding (parallel) sections (Leech 2012; see further Leech and Smith 2005). For specific background information to support my choices and decisions, I draw on a range of scholarly sources including:

- linguists with specialist experience of EModE drama and corpora (e.g. Crystal and Crystal 2005; Culpeper and Kytö 2010);

- literary critical studies, such as Braunmuller and Hattaway's (2003) edited volume; Dutton (e.g. 1991, 2000); Findlay (1999); Hunter (1997); Jardine (1983); Kinney (1999, 2002, 2009); Leggatt (1988, 1999); McRae (2003); and
- expert knowledge of specialists in Renaissance drama (Alison Findlay and Liz Oakley-Brown at Lancaster University).

My discussions in the next sections show that research in the literary critical tradition supports and adds value to a linguistic study of EModE drama, a point made in 1.2.

4.3.2.1 Issues and problems surrounding dates of the plays

Drama is known to have evolved during the course of the 16th and 17th centuries, both in style and in its surrounding socio-cultural context (see, for example, Culpeper and Kytö 2010:32-34; Hunter 1997; Leggatt 1988). In order to minimise the possible influences of changes in writing styles and/or language use over time on my results, it was essential for plays in the *NDC* to have been constructed between dates that approximate the period in which Shakespeare's First Folio plays were constructed (1589-1613; see 4.2.3 above). For parity, it was also important to balance the amount of dramatic dialogue written before and after 1600 with that in Shakespeare's First Folio. Shakespeare's style is argued as having changed around this time (by linguists, e.g. Crystal 2008:172; Murphy 2007:81-82, and by literary critics, e.g. Kermode 2000:13, 45-46), and it is likely that other dramatists' styles would also have evolved. Craig's (1999) corpus study shows that Ben Jonson's writing style changes over time, for instance.

However, I could not obtain sufficient digitised play-texts from the precise dates between which the plays in Shakespeare's First Folio were constructed to fulfil the requirements for each genre section of the *NDC* (bearing in mind that some were

unsuitable in other ways). The time-span of the plays in the *NDC* is therefore 1584-1626, a 42-year period beginning five years before the earliest Shakespeare play and ending thirteen years after the last one (bearing in mind that dating is not an exact science, as argued in 4.2.3). This is close to the time-span used by Craig (2011:61), who compares Shakespeare's plays to others dated between 1580-1619, arguing that "these four decades seem to be a reasonable span to represent the work of the immediately preceding generation, his direct contemporaries, and those who followed immediately afterward." I used plays from a slightly wider date band than Craig's, in order to fulfil my other inclusion criteria for the *NDC*. The 42-year period of the *NDC* approximates a single "generation" of language, which is considered to be a 40-year period in the construction of some diachronic historical corpora (notably the *CED*; see Culpeper and Kytö 2010:24-25, and the *Helsinki Corpus*; see Kytö 1996 [1991]).

The problems with dating EModE plays discussed in 4.2.3 mean that determining which ones actually fall within any target date band is open to some interpretation. The play-texts in the *NDC* are dated according to the year of construction, first performance, first printing or publication, depending upon the information available, according to the principles explained in 4.2.3. The digitised play-texts downloaded from *EEBO* all contain bibliographic information, including the year of publication. In some cases this may also be the year of first printing and/or performance, but this is not specified in the bibliographic details. Occasionally the date of first performance is incorporated into a full title of a play. Inevitably, the accuracy of some of the dates I use may be open to debate.

I now turn to the process of compiling the dramatic genre sections of the *NDC*, from available digitised play-texts in the target date band.

4.3.2.2 Issues and problems surrounding plays in each genre

The language in plays in different dramatic genres is known to vary, from evidence in existing studies. Hope states that:

> it is clear that certain registers of language are associated with certain genres (Renaissance pastoral attracts a relatively formal, archaic register, for example). Certain types of plot, and certain types of character, will entail certain types of vocabulary item – and there may even be syntactic expectations. (2010:171)

My (2007) research into key lexical bundles in Shakespeare's plays shows further evidence of variation in the kinds of formulaic language used relatively frequently by characters in different genres. For example, tragedies contain relatively more bundles which are part of *wh*-questions, and relatively fewer bundles which contribute to informational elaboration than the other two genres. I argue that this contributes to a dramatic atmosphere of suspense in tragedies (Demmen 2007:45-47). It is therefore important for the *NDC* to balance Shakespeare's First Folio with similar proportions of dialogue from comedy, history and tragedy plays, to avoid potentially skewing the results with language which is more typical of one genre than another. As noted with regard to Shakespeare's plays (in 4.2.3 above), the classification of the dramatic genres to which some contemporaneous plays belong is debatable. For this reason, although I consider the varieties of plays in each genre, I do not formally categorise them into sub-genres (e.g. romantic, city, pastoral and domestic comedies).

Other contemporaneous comedy plays were the most difficult to balance with those in Shakespeare's First Folio, since Shakespeare's comedy covers a number of different traditions, but does not extend to all that were popular around the same time. "City" comedies were popularised by dramatists such as Ben Jonson and Thomas Dekker (see e.g. McRae 2003:2), and feature London settings or London characters removed to other places. However, as Crystal and Crystal (2005:153) point out,

Shakespeare did not really write city comedies, the nearest qualifier being *The Merry Wives of Windsor* (see also Orlin 2003:159-161 and Twyning 2002:355). Shakespeare's comedies are set mainly outside London, and often in pastoral or rural settings. It was difficult to know whether or not a great quantity of city comedy would materially affect my results, since the distinction between city and non-city characters and settings is not clear-cut. Orlin (2003:160) argues that characters of high status and material wealth who are typically found in city comedies also feature in Shakespeare's plays, and Crystal and Crystal (2005:103) claim that Shakespeare's "country" or "rustic" characters "use styles of English not far removed from those of upper-class speakers". It therefore seemed prudent not to over-weight the *NDC* with city comedies, but also not to exclude them entirely, as they were so popular in the period under investigation in the study. The most prototypical city comedy in the *NDC* is Jonson's *Bartholomew Fair*, though the character types in his play *Volpone* (which is also included) potentially qualify it as such, despite its Venetian setting.

Pastoral comedies in the *NDC* include Fletcher's play *The Faithful Shepherdess*, and Lyly's *Gallathea* and *Alexander and Campaspe* (*Alexander and Campaspe* is compared by other scholars with Shakespeare's *As You Like It*, e.g. Dillon 2003:9; Shapiro 2002:318). Lyly's plays are limited to two in the *NDC*, because as Leggatt (1999:6, 12) notes, Lyly uses a particularly "extravagant" style and "mannered prose" known as "Euphuism". It was important not to over-represent any particular dramatist's language style in the *NDC*, again to avoid potentially skewing the results. Furthermore, Lyly is one of a number of playwrights who wrote for children's acting companies, i.e. those comprising boy choristers aged between about ten and twenty years (see e.g. Cerasano 2002:209-210; Munro 2005; Shapiro 2002). According to Shapiro (2002:315), Shakespeare wrote only for adult actors. This raised

the question of whether the age of the actors for whom a play was written would influence the language style of the dialogue and, consequently, whether to exclude plays for children's companies from the *NDC*. The discussions of Cerasano (2002) and Knutson (2002) suggest that drama written for children's companies was not censored in terms of content or style, nor was the dialogue tailored for younger speakers. Cerasano argues that:

> [...] boys performed all of the characters in the plays, including female parts. Like adult players, boys utilized a range of performance styles suitable to the roles they performed, and were capable of dealing with sophisticated rhetorical locutions. (2002:209).

Also, according to Leggatt (1999:8, 70-71, 136), some plays were performed by both adult companies and children's companies. It seems, therefore, that the dialogue was not materially different, though some types of characters might have been less convincingly portrayed by children, as suggested by Cerasano (2002:210) and Foakes (2003:28). On this basis, I do not exclude plays written for children's companies.

Romantic comedy is well represented in Shakespeare's First Folio and balanced in the *NDC*, for example by Heywood's *The Fair Maid of the West Part I* and the anonymous play *Mucedorus*. *Mucedorus* is possibly a contentious choice for inclusion, because it is in the so-called "apocrypha" of plays which have attracted claims of Shakespearean authorship or collaboration (according to Crystal and Crystal 2005:105; see further Hope 1994; Tucker Brooke 1908 and Wells et al. 1987). As indicated in 4.2.1, it is necessary to minimise the risk that any of the contents of the *NDC* might have been written by Shakespeare, in whole or in substantial part, because this would blur potential style distinctions between the dialogue in the two corpora. As far as possible, I include plays in the *NDC* for which the authorship is verifiable, but the evidence in a few cases is somewhat tenuous. To date, there is insufficient

evidence for *Mucedorus* to be widely accepted as authored by Shakespeare, but it is possible that this may change if new evidence comes to light in authorship attribution research. I include tragi-comedy in the *NDC* (Massinger's *The Bondman*), domestic comedy (Heywood's *How a Man May Chuse*), and comedy of humours (Chapman's *An Humerous Dayes Myrth*). Not all the other comedy plays fit neatly into particular sub-categories, and are included mainly on the basis of dates, to balance the amount of pre-1600 and post-1600 comedy with that in the *SDC*. The *NDC* comedies are listed in 4.3.3 (Table 6).

Shakespeare's history plays are all set in England, which constitutes the main criterion for those in the history section of the *NDC*. However, there were insufficient digitised English history play-texts within the target dates, so a few with non-English settings are also included. Some of these allude to England's relationship with other countries at the time, and feature English characters, for example Marlowe's *The Massacre at Paris*, Armin's *The Valiant Welshman* and Peele's *The Battle of Alcazar*. Marlowe's tragic history play *Tamburlaine Part I* is set far away from England in the old Ottoman Empire, but it is a popular and well-known play from within the target dates, and so is included. In deciding whether or not to include non-English history plays, I looked at samples to see whether the dialogue of English and non-English characters seems different in ways which would obviously influence my results (for example, through large quantities of dialect or the portrayal of foreign accents). There was very little evidence of this, and certainly no more than is in some plays with English settings that include non-English characters (e.g. the French princess Katherine in Shakespeare's *Henry V*). Accordingly, the inclusion of some non-English history plays was acceptable, as well as necessary to meet the dating and genre criteria.

The issue of the locations in the history plays brings me to the general question of whether locations in EModE plays (of any genre) influence the language used in characters' dialogue. Sullivan (2003:182-188) discusses views about the relevance or importance of geographical locations in Shakespeare's comedy plays, pointing out that the amount of detail given about them is sometimes quite minimal and occasionally inaccurate. This lends support to the idea that the locations themselves are less material than the creation of an alternative social space, away from England, in which different possible behaviours and courses of action become available to the characters. From this, it is more likely that the geographical setting would influence topical or localised language, rather than general style features in a character's dialogue. The corpus linguistic software settings are adjusted to minimise the occurrence of topical or localised results in my data, as explained in 3.4, making geographically-related language relatively unlikely to arise. Moreover, it would actually have been impossible to create a parallel corpus of plays with precisely similar locations to those in Shakespeare's plays, within the over-riding constraints of date and genre.

The evidence for the authorship of two of the history plays mentioned above is problematic, which I note in the interests of clarity and further explication of the inherent difficulties of studying historical texts. The author of *The Valiant Welshman* is tentatively accepted as being Robert Armin, based on the words "Written by R.A. GENT" in the earliest known quarto. *The Battle of Alcazar* is considered to have been written by George Peele based on an attribution of several quoted lines from the play in another contemporaneous publication, but no author's name appears on the title page of the earliest extant version, and it was not entered in the Stationers' Register. Since neither play seems to have been linked to Shakespeare (unlike those in the "apocrypha", mentioned above with regard to the comedy plays), they are included in

the *NDC*, even though their authorship remains open to challenge. The history plays in the *NDC* are listed in 4.3.3 (Table 7).

In addition to having varied geographical settings, as noted above, plays by Shakespeare and his peers were not always set in a contemporaneous historical period. For example, five of Shakespeare's tragedies feature classical settings (*Anthony and Cleopatra*, *Coriolanus*, *Julius Caesar*, *Timon of Athens* and *Titus Andronicus*). These are balanced to some extent, although not totally, by Jonson's play *Sejanus* and Marlowe's *Dido, Queen of Carthage* in the *NDC*. Revenge tragedy was popular in the Early Modern period, and is notably the theme of Shakespeare's *Hamlet* and *Titus Andronicus* (with other plays such as *Macbeth* including revenge elements as well, as argued by Boyce 1990:534). The *NDC* includes several popular contemporaneous revenge tragedies, such as Kyd's *The Spanish Tragedy*, Webster's *The Duchess of Malfi* and *The White Devil*, and Middleton's *Women Beware Women*. Shakespeare's *Romeo and Juliet* and *Othello* are tragedy plays with domestic, household and love elements, which are balanced in the *NDC* by Heywood's *A Woman Killed With Kindness* and the anonymous *Arden of Faversham*. The question of authorship arises again with *Arden of Faversham*, since it is argued by Kinney (2009:99) as being partly authored by Shakespeare. As with the comedy *Mucedorus*, mentioned above, it has not been accepted into the Shakespeare canon at the time of writing, and is therefore included in the *NDC*. The *NDC* tragedies are listed in 4.3.3 (Table 8).

Before leaving the subject of authorship, it is worth mentioning that although the number of Shakespeare's plays which are considered to be collaborations is minimised in the *SDC* (as stated in 4.2.1), collaborative plays are not excluded from the *NDC* if Shakespeare is not (at the time of writing) considered to be one of the authors. Beaumont and Fletcher's *The Maid's Tragedy* and *The Woman Hater* (a

comedy) are included, as is Middleton and Rowley's tragedy *The Changeling*, as they all usefully meet the dating and genre criteria. This causes no problem in my analyses, as my comparisons are not made at the level of individual authors in the *NDC*.

The amount of female dialogue in the comedy section of the *NDC* is a little low compared to the *SDC* (as quantified later on in 4.4), because of the limited availability of comedies of the appropriate dates and types. However, this is compensated for in the tragedy section, for which there was a greater choice of plays in the target date range containing relatively large quantities of female dialogue. Some of Shakespeare's tragedies, such as *Macbeth* and *Antony and Cleopatra*, feature relatively large roles for female characters, while others such as *Julius Caesar* and *Timon of Athens* are very much male-dominated. The mix of tragedies in the *NDC* is similar, with Marlowe's *Dido, Queen of Carthage*, Middleton and Rowley's *The Changeling* and Webster's *The White Devil* having female characters with relatively major roles and large amounts of dialogue, and others having very little (e.g. Marlowe's *Dr Faustus* and Jonson's *Sejanus*). To accumulate sufficient female dialogue to balance that in Shakespeare's First Folio, I did need to use a few plays which (unlike those by Shakespeare) have obviously female-oriented topics (*The Woman Hater*, *The Maid's Tragedy*, *A Woman Killed With Kindness* and *Women Beware Women*). This potentially skews the style of language in the plays, so I bear it in mind in considering results which appear to vary according to gender in my analyses.

This concludes my discussion of issues surrounding the selection of plays in the *NDC* to balance and represent the genres in Shakespeare's First Folio, in the wider context of popular varieties of EModE drama. The following further issues potentially impact on the comparability of other plays with those by Shakespeare:

- the popularity and success of the plays;
- plays performed in public and/or private theatres;
- the sex of the author(s);
- whether or not some plays were written for publication and/or performance;
- dialogue containing verse and prose; and
- characters who cross-dress between genders.

In the next section, I discuss each of the above points briefly.

4.3.2.3 Further issues, problems and questions surrounding the choice of other contemporaneous plays

Shakespeare was a successful playwright, so it is sensible for a parallel corpus to comprise works that were also generally popular or successful (either at the time or through subsequent revival). This is because variation in the style of language in Shakespeare's plays and that in much less successful or less popular works might simply be attributable to a difference in the quality of the dramatists' writing. For more detailed discussions of the relative popularity of Shakespeare's contemporaries, see for example Leggatt (1988:167-186).

As far as possible, the *NDC* includes plays which were, like Shakespeare's, performed at public theatres and also at court, but not in other private settings. According to Crystal and Crystal (2005:7, 63, 181), Shakespeare's plays were performed at the Globe (a public theatre), and also privately at court for the monarch and other elite members of society, especially after Shakespeare's acting company had obtained the patronage of King James I. Dutton (2011) argues that plays which were performed at court benefited from an added air of glamour that would have increased their appeal to paying customers at public theatres. There is no information about whether the edition of Shakespeare's plays used for the *SDC* or the editions of plays on

EEBO are versions written for court or public performance (other than what can be gleaned from the prologues or other preamble, where this exists). However, according to Dutton (2011), public performances were subject to limited time constraints, whereas private performances could go on for many hours. This makes it more likely that the Shakespearean play-texts in the *SDC* are the court versions, because of their relatively long length compared to other plays (evident in the comparison of word counts for the corpora in 4.4 below). That would also apply to other relatively long plays in the *NDC* (such as Jonson's *Bartholomew Fair*; see the word counts in Table 6 in 4.3.3 below). As I argued in 4.2.3 that dialogue in early editions was added, deleted or revised to suit different audiences (Greenblatt 1997:67; Petersen 2010:13), this raises the question of how far it might influence language styles in the play-texts.

Dutton (2011) points out that making detailed revisions for different audiences would have added to the cost of putting the plays on, and Crystal and Crystal (2005:39) argue that sections of plays were simply left out by the acting companies that performed them, according to the requirements of the theatres in which they were appearing. These arguments suggest that although sections of dialogue were cut or added, particular authorial style features are much less likely to have been altered for private and public theatregoers. There is also evidence to suggest that the orientation toward different audiences is located more in the prologues than in the main body of the play. For example, the earliest extant edition of Marlowe's *The Jew of Malta* includes two prologues: "The Prologue Spoken at Court" and "The Prologue to the Stage at the Cock-Pit" (a public theatre). The former is much more humble and deferential, though the text of the play itself is the same. Because of this, and other constraints on the selection of play-texts, I did not pursue the distinction between court or public versions further. It seems unlikely to have any bearing on the high-

frequency, non-localised language features in my results. For more discussion on the private and public performance of plays, see H. Berry (2002); Butler (2003); Cook (1997:308, 319-320); Foakes (2003).

As indicated above, private drama which was performed at venues other than court, e.g. in private homes for small, invited audiences, is excluded from the *NDC*. Unlike plays performed at court and in public theatres, private drama was not subject to the regulation and approval of the Master of the Revels, an official who had the authority to insist on alterations to the language or the plot of a play, if he considered them unsuitable for the public good or offensive to the monarch and court members (see further Crystal and Crystal 2005:62; Dutton 1991, 2000). Since the authors of private drama in non-court settings could avoid this censorship, the language styles in their works are potentially more liberal than those in public or court drama (which might influence my results).

The decision to exclude private drama outside of that performed at court automatically excludes plays written by and/or acted by women, who only wrote or performed in private drama in the late 16th and early 17th centuries (see McRae 2003:7; Westfall 2002:274; see also Findlay 1999:7-8, 114). Public drama was written by men and acted by men and boys (Findlay 1999:1; see also H. Berry 2002:148 and Orgel 1996:1-9). Including only male-authored plays in the *NDC* maintains consistency with Shakespeare's plays, and avoids the introduction of another variable which potentially influences language styles in drama: authorial gender. At a time when women had very unequal status in law and in public life (as argued by, e.g., Nevalainen and Raumolin-Brunberg 2003:113-115), the construction of language styles of men and women in plays by male and female dramatists is interesting, although outside the scope of my study. Male authorship of the three anonymous plays

in the *NDC* cannot be verified, but is reasonable to assume from their performance in public theatres and/or from discussions of putative authorship (e.g. Kinney 2009; Tucker Brooke 1908).

Dutton (2000:109) considers whether the relatively long length of plays written by Shakespeare and Jonson, noted above, indicates that they were written with readers in mind, instead of a theatre audience. This potential difference in orientation has led to a distinction between "literary" and "non-literary" dramatists in some circles, so I investigated whether it would be likely to influence language styles to an extent that might affect my results. According to McRae (2003:7), most playwrights in the Early Modern period did not seek publication, but rather retained the advantages of control over their own plays, and Greenblatt (1997:67-68) claims that Ben Jonson was unusual for publishing a folio of his own works. However, Erne (2003) argues that Shakespeare's theatre company did seek to have plays published a year or two after they were first performed, and that excerpts from Shakespeare's plays were published in other contemporaneous books. Findlay (personal communication, 18.12.09) suggests that the goals of Jonson and his contemporaries were similar, i.e. to attract paying theatre audiences to their plays, but that Jonson was simply a little ahead of his contemporaries in his ambitions for publication and in the manner he distributed his work. I could find no evidence that an orientation to publication or performance would affect language styles in the plays, so Jonson's work is represented in the *NDC*. He is one of the most well-known and popular of Shakespeare's contemporaries, and his works are useful for comparison (as indicated in 4.3.2 above).

Finally, as with Shakespeare's plays, those in the *NDC* feature dialogue comprised of verse as well as prose (see Crystal and Crystal 2005:165 for a breakdown of the verse and prose lines in each of Shakespeare's plays; see also Crystal 2008:210-

219). This inevitably has some influence on the language styles of Shakespeare and other dramatists, but does not bias either corpus. Also, characters who cross-dress (between genders) feature in both corpora (e.g. in Shakespeare's comedies *As You Like It* and *The Merchant of Venice*, and Heywood's comedy *The Fair Maid of the West Part I*). Again, although this influences characters' language styles (see further e.g. Findlay 1999; Rackin 2003:114), neither corpus is biased. Dialogue spoken by and to cross-dressed characters represents only a small proportion of the contents of the corpora, and is unlikely to affect the high-frequency results underlying my analyses.

In this section, I have shown what a complex task it is to construct a corpus that is similar in size and comparable in content to Shakespeare's First Folio, and I have clarified the principles I followed and the limitations faced. I now list the plays which comprise the *NDC*.

4.3.3 List of plays in the *Non-Shakespearean Early Modern English Drama Corpus*

The plays in the *NDC* are displayed in Tables 6, 7 and 8 (comedies, histories and tragedies, respectively). They are listed in chronological order (according to date of construction, first performance, first printing or first publication), and word counts are also given (from *WordSmith*). The total word count for the *NDC* is 796,582.

Table 6. Comedy plays in the *NDC*

Author	**Title**	**Date of construction, first performance, printing or publication**	**Date of edition in the corpus**	**Total dialogue word count**
John Lyly	*Alexander and Campaspe*	1584	1584	12,214
John Lyly	*Gallathea*	1588	1592	13,140
Robert Greene	*Friar Bacon and Friar Bungay*	1589	1594	15,893
George Peele	*The Old Wives Tale*	1595	1595	7,584
George Chapman	*The Blind Beggar of Alexandria*	1596	1598	13,099
Thomas Heywood	*The Fair Maid of the West Part I*	c.1597-1603	1631	14,687
George Chapman	*An Humerous Dayes Myrth*	1597-1599	1599	15,872
Henry Porter	*The Two Angry Women of Abington*	c.1598	1599	25,660
Anonymous	*Mucedorus*	1598	1598	10,846
Thomas Dekker	*Old Fortunatas*	1599	1600	22,255
Thomas Heywood	*How a Man May Chuse*	1602	1602	21,324
Ben Jonson	*Volpone*	1606	1616	26,243
Francis Beaumont and John Fletcher	*The Woman Hater*	c. 1606	1607	22,313
George Wilkins	*The Miseries of Inforst Marriage*	1607	1616	23,778
John Fletcher	*The Faithful Shepherdess*	c. 1608-1609	1610	19,892
Ben Jonson	*Bartholomew Fayre*	1614	1631	34,236
Philip Massinger	*The Bondman*	1624	1624	19,945
Total non-Shakespearean comedy dialogue				318,981

Table 7. History plays in the *NDC*

Author	Title	Date of first construction, performance, printing or publication	Date of edition in the corpus	Total dialogue word count
Robert Greene	*The Scottish History of James the Fourth*	1590	1598	19,777
Christopher Marlowe	*Tamburlaine Part I*	c. 1590	1633	14,975
Christopher Marlowe	*Edward II*	1592	1604	20,617
George Peele	*The Famous Chronicle of Edward I*	1593	1593	21,353
Christopher Marlowe	*The Massacre at Paris*	1593	1594	9,691
George Peele	*The Battle of Alcazar*	1594	1594	9,594
Anthony Munday	*The Death of Robert Earl of Huntingdon*	c.1598	1601	22,296
Thomas Heywood	*Edward IV Part I*	c.1599	1600	22,369
Thomas Heywood	*Edward IV Part II*	c.1599	1600	24,025
Anonymous	*The Life of Sir John Oldcastle*	c. 1599	1600	20,985
Thomas Heywood	*If You Know Not Me, You Know Nobody Part I*	1605-1606	1607	11,366
Thomas Dekker	*Sir Thomas Wyatt*	1607	1607	11,013
Robert Armin	*The Valiant Welshman*	c. 1610-1615	1615	15,876
Thomas Drue	*The Duchess of Suffolk*	1624	1640	16,352
Total non-Shakespearean history dialogue				240,289

Table 8. Tragedy plays in the *NDC*

Author	Title	Date of first construction, performance, printing or publication	Date of edition in the corpus	Total dialogue word count
Thomas Kyd	*The Spanish Tragedy*	1587	1592	20,853
Christopher Marlowe	*The Jew of Malta*	1589-1590	1633	17,858
Anonymous	*Arden of Feversham*	1591	1633	19,759
Christopher Marlowe	*Dr Faustus*	1592	1604	11,429
Christopher Marlowe	*Dido, Queen of Carthage*	1594	1594	13,578
Thomas Heywood	*A Woman Killed With Kindness*	1603	1607	16,191
Ben Jonson	*Sejanus*	1603	1616	26,091
Francis Beaumont and John Fletcher	*The Maid's Tragedy*	c. 1610	1619	20,994
John Webster	*The White Devil*	1612	1612	24,461
John Webster	*The Duchess of Malfi*	1614	1623	22,811
Thomas Middleton and William Rowley	*The Changeling*	1622	1631	18,257
Thomas Middleton	*Women Beware Women*	c. 1612-1627	1657	25,030
Total non-Shakespearean tragedy dialogue				237,312

The authorial styles of 23 different playwrights are represented in the *NDC* (assuming that the anonymous plays were written by different individuals, which is unverifiable).

The size and content of the *SDC* and the *NDC*, detailed in this and the preceding sections, have important implications for the methods of statistical comparisons between the two corpora, which lead to the identification of similarities and differences in the language styles of Shakespeare and the other contemporaneous dramatists in my analyses in chapters 6 to 8. Therefore, given the necessity for choices and compromises detailed above in this chapter, it is now helpful to summarise the contents of both corpora quantitatively, and highlight any shortfalls or imbalances in the parallel sections which might influence language styles. In 4.4, I present and discuss a side-by-side comparison of word counts for the main components into which the two corpora can be broken down.

4.4 Quantitative comparison of the contents of the *Shakespearean Drama Corpus* and the *Non-Shakespearean Early Modern English Drama Corpus*

Table 9 on the next page gives the word counts of the two corpora overall, and broken down by genre, sex of speaking character and date (pre-1600 or post-1600). More detailed breakdowns of the components of the corpora are given in Appendix II.

Table 9. Comparison of the size and structure of the *SDC* and the *NDC*

	SDC	*NDC*
Overall		
All characters in the corpus	797,054	796,582
Average length of play-text	22,140	18,525
Breakdown by genre		
All comedy characters	330,616	318,981
All history characters	235,771	240,289
All tragedy characters	230,667	237,312
Breakdown by date		
All characters in pre-1600 plays	429,421	420,409
All characters in post-1600 plays	367,633	376,173
Breakdown by sex of character		
All female characters	140,227	153,433
All male characters	656,319	640,139
All characters of unknown sex	459	2,375
Both sexes speaking in unison	49	635
Word counts from Scott's (1999) *WordSmith Tools* V.3.0		

Table 9 shows that when the corpora are broken down into component parts, the discrepancy between dramatic genres is greatest between the comedy sections, in which the *SDC* is larger than the *NDC* by 11,635 words. The history and tragedy sections of the *NDC* are correspondingly larger than those in the *SDC* (by 4,518 and 6,645 words, respectively). The *SDC* contains just over 9,000 more words from plays written before 1600 than does the *NDC* (and just over 8,500 fewer words of dialogue written after 1600). It was quite difficult to find tragedy plays dated before 1600 and history plays dated after 1600 which were closely comparable to those by Shakespeare. Overall though, as Table 9 shows, the sizes of the *SDC* and the *NDC* are extremely close (differing by fewer than 500 words). Bearing this in mind, together with the other constraints highlighted in 4.2 and 4.3, the (substantial) efforts to construct a corpus of play-texts which would approximate the *SDC* in size, date, content and genre have been mainly successful.

I touched briefly on problems with balancing the amounts of male and female dialogue, and the relative dominance of male characters, in 4.3.2.2 above. Table 9

shows that the *SDC* contains just over 16,000 more words of male dialogue than the *NDC*, and just under 13,000 fewer words of female dialogue. As indicated at the bottom of Table 9, small amounts of dialogue in each corpus are spoken either by characters whose sex could not be determined (from descriptions in lists of dramatis personae, according to vocatives, pronouns or other forms of address, or from the context), or by male and female characters speaking in unison.

Table 9 also reveals that on average Shakespeare's plays are 3,615 words longer than those by his contemporaries. As suggested in 4.3.2.3, this might be because they are the "uncut" versions, i.e. longer versions of plays for performance at court, or perhaps for an audience of readers rather than theatregoers. 43 play-texts in the *NDC* approximate the same amount of dialogue in the 36 play-texts of the *SDC*. Word counts for each play in the *SDC* (in 4.2.3) and the *NDC* (in 4.3.3) show that the length ranges between 14,630 (*A Comedy of Errors*) and 29,832 (*Hamlet*)[38] in the *SDC*, and 7,584 words (Peele's *The Old Wives Tale*) and 34,236 (Jonson's *Bartholomew Fair*) in the *NDC*. Having compared the amounts of dialogue in different sections of corpora, I now look briefly at the numbers of characters.

Table 10 gives the number of male and female characters in each corpus, in total and broken down by genre, together with the number of characters whose sex could not be determined.

Table 10. Number of male and female characters in the *SDC* and the *NDC*, by genre

	SDC				***NDC***			
	Male	Female	Unknown	All	Male	Female	Unknown	All
Comedy	322	72	1	395	348	84	13	445
History	421	98	0	519	519	64	1	584
Tragedy	367	39	3	409	285	53	13	351
All	1,110	209	4	1,323	1,152	201	27	1,380

[38] Crystal and Crystal (2005) also identify these as the shortest and longest plays in their database of the plays, with not dissimilar word counts (14,415 and 29,844, respectively).

Table 10 shows that in both corpora there are many more male characters than female (1,110 to 209 in the *SDC*, and 1,152 to 201 in the *NDC*). Overall, both corpora feature similar proportions of male and female characters, i.e. between five and six times as many men to women. Looking at the figures in Table 10 by genre, in the *SDC* the male-to-female character ratio is similar in comedies and histories (about 4.5 to 1 and just over 4 to 1, respectively), but much higher in tragedies (over 9 to 1). In the *NDC*, the ratio of males to females is not dissimilar to the *SDC* in comedies (just over 4 to 1), but double that of the *SDC* in histories (at just over 8 to 1) and about half that of the *SDC* in tragedies (at just over 5 to 1). This is because I included plays with relatively more female dialogue in the tragedy section of the *NDC* in order to make up a shortfall in the comedy section, as explained in 4.3.2.2.

The plays in the *NDC* are similar to Shakespeare's plays in featuring fewer female characters than male and much more male dialogue than female dialogue. The exception is Lyly's *Gallathea*, which has very slightly more female dialogue. It is tempting to jump to the conclusion that this reflects the lower importance of women in patriarchal Early Modern society, but that would seem to be somewhat at odds with the fact that women were part of the (paying) theatregoing audiences at the time. Findlay (1999:6) argues that women were "an audience of customers whose situations, opinions and tastes male dramatists probably responded to". Commenting on Shakespeare's plays, Crystal and Crystal (2005:135) suggest that the comparatively low ratio of female to male characters is perhaps due to practicalities, since actors were adult men, apart from a few boys who played the female roles. The storylines following the actions of royal and aristocratic men in the history plays would explain the relatively low presence of women and female dialogue in that genre, at least, though there are of course some major female roles in the histories (e.g. Queen

Margaret, who appears in the three parts of Shakespeare's *Henry VI*, and in *Richard III*).

Having documented and discussed the decisions which underpin the compilation of the *NDC*, in relation to the contents of the *SDC*, I give a brief concluding summary in the next section. I then move on to the ways the play-texts in the *NDC* have been treated to make them more suitable for analysis with corpus linguistic software tools, in chapter 5.

4.5 Summary and conclusions

In this chapter I have explained that the plays in the *SDC* are from an existing Shakespeare corpus which I obtained (from Mike Scott), adapted and prepared for previous research (in 4.2). Since this corpus is already annotated and suitable for the methods applied in this study, using it afforded sufficient resources to meet the research aim of producing a new corpus of other contemporaneous EModE plays. There is an issue of compatibility between modern and historical source texts, as discussed in 4.2.2, but this is smoothed out to a great extent through spelling regularisation (discussed later on in 5.4), and does not cause any major problems.

I have also argued that it was justifiable to construct a new, highly specialised parallel corpus for Shakespeare's First Folio, and I have explained and illustrated how the careful choice of play-texts for the *NDC* means it is as balanced, representative and comparable as possible, within the main constraints of dating, genre and availability (in 4.3). The selection process for the *NDC* was supported by scholarly commentary and critical editions of the plays, and the final choice takes into account the wider landscape of EModE drama as well as the nature of Shakespeare's plays. The decisions made have enabled me to construct a corpus which is large enough, and which

includes comparable amounts of popular comedy, tragedy and history dialogue from before and after 1600.

Like the *SDC* (and other Shakespeare corpora), the *NDC* is not without its "problem plays", due to gaps in historical evidence about dating and questions of authorship (e.g. *Mucedorus* and *Arden of Faversham*, discussed in 4.3.2.2). In 4.2 and 4.3 I have demonstrated that dating and authorship are two difficult issues which will inevitably face anyone wishing to investigate EModE drama (and other types of historical texts), hence my exhaustive discussions to make clear how I handled them, and to clarify any potential impact on my results. There may well be arguments for different compilations of other contemporaneous plays as a parallel corpus to compare with Shakespeare's First Folio, but in this chapter I have explained how and why that in the *NDC* serves the particular aims and methods of my study.

Whilst the *NDC* serves the primary purpose of a reference corpus for Shakespeare's plays in this study, and enables the answering of research question 1 (concerning similarities and differences between the authorial styles of Shakespeare and a range of his peers), it also provides scope for much future research. Since it is specifically designed for comparison with Shakespeare's First Folio, it is not entirely representative of EModE plays in general. As I point out with regard to future research possibilities in my conclusions in 9.4, however, the *NDC* could be enlarged to become a more general EModE drama corpus.

CHAPTER 5. PREPARING EARLY MODERN ENGLISH PLAY-TEXTS FOR THE APPLICATION OF CORPUS LINGUISTIC METHODS

5.1 Introduction

This chapter focuses on the processes applied to the play-texts in the *NDC*, post-compilation, and addresses research questions 3.3-3.4 (concerning further preparation of corpus texts and the issue of historical spelling variation). I report my experiences with some relatively new software tools used to prepare the corpus texts, and I discuss the benefits they offer (particularly to researchers working with historical data). Three processes improved the prospects for generating useful results with the methods discussed in chapters 2 and 3. Firstly, the quantitative results from the corpora needed to be based on dialogic text only, and in 5.2 I explain how non-dialogic text was excluded through annotation. Here, I discuss my experiences with using the scripting language PHP to increase the efficiency of annotation. This is a relatively new technique which requires specialist knowledge, but offers potential benefits for corpus linguists, and is a topic of ongoing research at Lancaster University (by Andrew Hardie)[39].

Secondly, missing text and textual anomalies in the digitised *EEBO* play-text files required assessment and some correction by comparison with the corresponding printed manuscript facsimiles on *EEBO*. I discuss this in 5.3. Transcription methods and levels of accuracy vary, and it is important to check digitised files for accuracy (as is also done, for example, by McIntyre and Walker 2011:110 in their corpus study of digitised EModE texts from a range of sources). My findings are of potential interest to other researchers who may wish to make use of the *EEBO* digitised play-texts.

Thirdly, non-standardised spelling in the early extant EModE play-texts in the *NDC* needed to be regularised, in order to improve precision and recall when the

[39] See http://www.ling.lancs.ac.uk/activities/970/ (last accessed 13.07.12).

results were generated, as I discuss in 5.4. Spelling variation reduces the ability of corpus linguistic tools which work by orthographic matching to retrieve results, since frequency counts are split over the different variants of a single word. For example, the research of Archer et al. (2003), Baron et al. (2009) and Rayson et al. (2007) shows that the effective identification of key results using corpus linguistic methods is reduced by variation in spelling. The spelling in the Shakespearean play-texts which now comprise the *SDC* was regularised using *VARD 2* (Baron and Rayson 2008) (version 2.1.5) for my 2009 research, although the texts were already modernised to some extent (as discussed in 4.2.2). Over 2,000 spelling variants were regularised, and my findings supported Rayson et al.'s (2007:2) claim that even modernised EModE texts benefit from regularisation when undertaking corpus research. I do not discuss this further in 5.4, since it is covered in Demmen (2009:62-66). Instead, I concentrate on the treatment of the *NDC* texts using a newer version of *VARD 2* (version 2.3), which has not been applied in many studies thus far. Other spelling regularisation tools exist, e.g. "Intelligent Archive" (see Craig and Whipp 2010), used by Craig and Kinney (2009:xviii) with their EModE drama corpus (mentioned in 4.3.1), and ZENSPELL, used by Schneider (2002) with 18th-century English. However, my discussions will be confined to *VARD 2*.

5.2 Annotation and encoding of the play-texts in the *Non-Shakespearean Early Modern English Drama Corpus*

In addition to excluding non-dialogic text from computations made by the corpus tools, mentioned in the previous section, annotation also preserves information such as stage directions and speaker labels in the texts. This helps make sense of what is going on in the dialogue and aids the interpretation of results during analysis. Additionally, annotation enables some useful meta-data about the contents of each play-text file in

the corpus to be encoded for reference. To some extent, annotation can be automated by searching and replacing text to be tagged using regular expressions. With a text editor such as *Notepad++*[40] this can be carried out on multiple texts simultaneously. Additional manual fine-tuning is usually required to tag individual items which cannot be globally searched. The annotation process is therefore time consuming and open to human error. This is pointed out by Baker (2006:42), who also emphasises that the direct benefit of annotation to the interpretation of results in a study needs to be assessed carefully at the outset. The annotation of the *NDC* is therefore limited to what was essential to serve the needs of the present study, but I use a conventional system which has potential for exploitation in future research, and to which further degrees of annotation can be added later if necessary.

My rationale for annotating and marking up the *NDC* play-texts is informed by those of other scholars who have built corpora containing EModE drama, e.g. Archer and Culpeper (2003), Culpeper and Kytö (2010), Kytö and Walker (2006) and Lutzky (2009a, 2009b and 2012). The annotation I use is encoded in XML (eXtensible Mark-up Language) tags. XML tags are based on a standardised coding system for electronic texts, but their contents can be customised (Baker 2006:38-42). They are therefore useful in corpus linguistics, because they can be tailored to different kinds of texts whilst still being interpretable by a range of other computer programmes that may be required for analysing them. As noted in 3.3.2, *Wmatrix* will only work with well-formed XML tags. *WordSmith* excludes from computation any text bounded by a pair of angle brackets. This automatically includes XML tags (which are bounded by pairs of angle brackets), although, as mentioned in 3.3.2, it inexplicably picked up one word (*who*) from the speaker identity tags (which was fortunately an isolated anomaly).

[40] Currently free to download. See http://notepad-plus-plus.org/ (last accessed 31.08.12).

Devising annotation systems is not particularly easy, and compatibility with different kinds of corpus linguistic software tools is an issue.

Andrew Hardie (in preparation, and personal communication, 05.05.10) argues that the scripting language PHP can be used to increase the speed and efficiency of annotation. It requires a reasonable level of knowledge of computer programming and regular expressions, however, which I could not have acquired sufficiently in the time available for the project. Therefore, he and I discussed the most advantageous ways to annotate the texts, and he wrote the PHP scripts which I could then execute and adapt in minor ways (e.g. by altering elements of the regular expressions in the scripts in order to change the search-and-replace parameters). PHP requires a text editor in which to write the scripts, and we used *Notepad++*. Among its useful features (which I discuss a little further in 5.5) is the display of XML tags in different colours from the main body of the text. PHP scripts automated the tagging of the speaker identities of over 31,000 speech turns in the *NDC*, as I explain in more detail below. The other tag-types in the corpora were not in sufficiently standard forms to be annotated automatically.

The XML tags used to mark up the contents of the play-texts in the *NDC* are summarised in Table 11 below, which is followed by a brief explanation of each one.

Table 11. Encoding conventions used in the *NDC* in the form of XML tags

<text id=""> </text>	short code for identifying the play title and genre, with an end marker showing where the zone of the play-text finishes
<ref c=""/>	reference and bibliographic information about the play-texts
<frontmatter c=""/>	additional text preceding the dialogue of the play
<endmatter c=""/>	additional text following the dialogue of the play
<comment c=""/>	typewritten notes and/or markers highlighting an anomaly or problem in the dialogic text, e.g. missing or unclear words
<stagedir c=""/>	stage directions
<sceneid=""> </scene>	start and end tags marking acts and scenes
<u who="">	speaker identification tags

Short code for identifying the play title and genre

Each play in the *NDC* has a short title code comprising an initial letter N (denoting non-Shakespearean plays), a second letter identifying the genre (C for comedy, H for history and T for tragedy), and then a short form or acronym of the title of the play. For example, *The Duchess of Malfi* is coded as NTDOM, and *Bartholomew Fair* as NCBFAIR. These are inserted as text identification ("id") tags at the top of each play-text file, and the end marker inserted after the final line, to serve as boundaries when joining multiple files. Play-text files were labelled using the text-id, for consistency. Play-texts in the *SDC* are labelled in a similar format, but with the initial letter S, e.g. SCMWW for *The Merry Wives of Windsor*. Text-ids for all the plays in both corpora are shown in Appendix IV.

Reference and bibliographic information about the play-text

Baker (2006:40) explains that including tagged "headers" enables the retention or inclusion of "meta-linguistic" information about the texts in a corpus. This is useful for reference. The digitised play-texts on *EEBO* already contain information such as the date and bibliographic name or number, the extended title, author's name and lifespan, and date of publication. I encoded all this in a single "reference" tag, to form a header in each play-text. The header information could be broken down into separate components, as is the case in the *CED* text files, if there was a need to search on individual components such as date. However, this is not necessary in my study. The header tag from the top of the play-text of Webster's tragedy *The White Devil*[41] is shown on the next page.

[41] Non-dialogic text in the corpora, including the header tags, is not subjected to the spelling regularisation process discussed in 5.4. In the text of the thesis, I standardise the titles of the plays to the modern forms by which they are commonly known today, for convenience and brevity.

```
<ref c="Author: Webster, John, 1580?-1625?
Title: The white diuel, or, The tragedy of Paulo Giordano
Vrsini, Duke of Brachiano with the life and death of Vittoria
Corombona the famous Venetian curtizan. Acted by the Queenes
Maiesties Seruants. Written by Iohn Webster.
Date: 1612
Bibliographic name / number: STC (2nd ed.) / 25178
Bibliographic name / number: Greg, I, 306(a). /
Physical description: [88] p.
Copy from: Bodleian Library
Reel position: STC / 1296:01"/>
```

Other non-dialogic text preceding and following the content of the play

Some play-texts feature an introductory preamble for the benefit of the players or the audience, before the dialogue of the characters begins. This typically includes prologues, dedications to the dramatists' patrons or friends, and/or a list of dramatis personae. It is all encoded in a single <frontmatter c=""/> tag in each play-text. Any text which comes after the dialogue of the play ends (typically an epilogue, or the printer's details) is encoded in a single <endmatter c=""/> tag.

Comments

Missing or unclear text in the play-text files is indicated by three leader dots between square brackets: [...], a convention already in place in some of the digitised files downloaded from *EEBO*. The marker is encoded in a comment tag, as in the following line from Marlowe's tragedy *The Massacre at Paris*:

```
And made <comment c="[...]"/> look with terror on the world:
```

Apart from the above "missing" marker, a few other brief notes are encoded in comment tags in the play-texts, such as places where a speaker's identity is unclear.

Stage directions

Stage directions are marked off in stage direction tags, e.g.:

```
<stagedir c="Exeunt."/>
```

Acts and scenes

As noted in 1.5.2, not all early extant EModE play-texts are divided up into acts and scenes. In the *NDC*, those which are are marked off into zones between a scene-id tag containing the text-id code and an end-of-scene tag. For example, Act I, scene i in *Tamburlaine Part I* is bounded by the following tags:

```
<sceneid="NHTAM_I_i">
</scene>
```

Speaker identification tags

Each speech turn in the play-texts downloaded from *EEBO* is preceded by a speaker label, in most cases an abbreviation of the character's name followed by a full-stop. These speaker labels needed to be converted to speaker-id tags. As mentioned at the start of this section (5.2), the annotation of speaker-id tags was automated to a great extent using PHP scripts. These are given in Appendix V. A single PHP script (written by Andrew Hardie) carried out the following set of commands:

(i) the identification of speaker labels, by searching for a single word followed by a full-stop on a line by itself, e.g.:

```
Lodovico.
```

(ii) the insertion of a tag containing that character's name immediately after the speaker label (a "u who" tag from Table 11 above), e.g.:

```
<u who="Lodovico">
```

(iii) the insertion of an end of utterance tag </u> before each speaker label, to mark off the end of the previous character's speech.

Lodovico's speech turn was then marked off by the end-of-utterance tag preceding the next speaker's turn, which was prompted by PHP finding the next speaker label in the play-text. The end-of-utterance tag before the first speaker label in each play-text was

redundant and was deleted manually, and a final end-of-utterance tag was added manually after the last speech turn in the play-text. In just a few seconds, the execution of this single PHP script annotated the vast majority of 31,000 speaking turns in all 43 play-texts in the corpus (it also inserted the text-id tags, discussed above, and saved the annotated text files as new XML files).

Often, however, the speaker labels for a single character are not consistent in EModE play-texts, because of non-standard spelling and abbreviations. This variation caused PHP to code them with separate "u who" tags (meaning that a single character's speech was split across several different speaker-id tags). That would have made it more difficult to extract all the dialogue of a single character, which was desirable for creating separate male and female data files later on (mentioned below). To address the problem, I used a second PHP script which identified and listed all the variants of "u who" tags in a play-text. I could then identify potential variants of a single speaker label, verify them as belonging to one character in the play-text, and convert multiple variants to a standard speaker-id tag using a third PHP script. For example, in the play-text of *The Duchess of Malfi* downloaded from *EEBO*, speaker labels for the character Ferdinand were variously abbreviated to "Fer.", "Ferd.", "Fred.", "Ford." or "Berd.". Each variation was initially assigned a different "u who" tag by the first PHP script. These were then identified with the second script, and finally replaced with a standard tag: <u who="Ferdinand"> by the third script.

Following the automated annotation of speaker-id tags as explained above, a relatively small amount of manual fine-tuning was necessary to correct text which fitted the search parameters of the first PHP script, but which did not constitute speaker labels. These were instances of single words of dialogue followed by a full-stop (i.e. one-word speech turns, such as "Good."). I could only do this by scrutinising

the corpus texts and checking them line by line, which also enabled me to pick up a few non-standard speaker labels that the first PHP script could not capture. These were speaker labels **not** followed by a full-stop, which were rare in most of the *NDC* play-texts, although prevalent in a few. In these cases, the "u who" tag and end-of-utterance markers had to be inserted manually. The original speaker labels in the play-text files also had to be encoded between pairs of angled brackets to isolate them from the rest of the dialogic text, a process which in retrospect could have been included in the first PHP script. It was quick and easy to go back and make global replacements, however, whilst I was checking the corpus texts and carrying out the manual annotation of other tag-types discussed above (using *Notepad++*).

Following the annotation of the speaker-id tags in each successive play-text, I used a fourth PHP script to count the number of words of dialogic text, i.e. everything contained between "u who" tags and end-of-utterance tags (apart from anything marked with other tags, e.g. comments and stage directions). PHP defines a word as a "string containing alphabetic characters, which also may contain, but not start with "'" and "-" characters"[42], and its word count function produces results that are not entirely consistent with those from other programmes such as *WordSmith*. However, it provided a quick guide to the amount of dialogue harvested from each play-text, which was useful in building up each section of the *NDC* to a size approximating that in the *SDC*.

I constructed a spreadsheet logging the name and sex of each character in every play during the annotation process, to facilitate the rapid block extraction of male and female dialogue with a fifth PHP script. This enabled me to create separate components of the corpus for analysis of selected results by gender, which is

[42] See http://uk3.php.net/manual/en/function.str-word-count.php (last accessed 10.08.12).

occasionally useful in the present study (e.g. in 8.3 for the word cluster I PRAY YOU), and which will benefit future research into language styles and gender in the plays (which I suggest as a useful direction in 9.4).

Other than the annotation explained above, and some correction of gaps and mis-transcribed text in the *EEBO* digitised text files to increase their accuracy (discussed in the next section), nothing was added to the play-texts in the *NDC*. Nothing was deleted apart from extra blank lines, spaces and unusual characters such as hash signs # which might interfere with the orthographic matching processes of the corpus analysis software, following Lutzky (2009b:1). The only characters which do not stand for themselves in the play-texts are the angle brackets < and > which surround the encoded information. Following Kytö and Walker (2006:37-38), I did not alter the lineation of the play-texts, but I removed hyphens from words which were split at the line break for the printers' convenience. This is because they would otherwise artificially inflate word counts for the *NDC* texts (compared to the *SDC* texts, which do not have words split at line breaks).

5.3 Missing text and other transcription issues in the digitised play-texts

As mentioned in the previous section (with regard to comment tags), some of the digitised *EEBO* play-texts already contain markers indicating that text is missing. The amount of missing text represented by these markers needed assessing in my study, because the *NDC* is intended to be a parallel corpus for Shakespeare's First Folio comprised of complete play-texts. Play-texts with substantial gaps would have been less suitable for inclusion. I checked the marked gaps against the corresponding facsimile printed manuscript files of the play-texts on *EEBO*. In most cases, the gaps corresponded to faint manuscript print, and where I could read this reliably I filled in

the gaps in my versions of the play-texts. Whilst comparing the digitised and printed manuscript files closely in an effort to fill the marked gaps, I also detected some unmarked gaps and some inaccurately-transcribed spellings. A few play-texts were particularly badly affected, with anomalies in nearly every line (notably Heywood's *Edward IV Part I* and Marlowe's *Tamburlaine Part I*). This finding was surprising, because an "established standard of 99.995% character accuracy" is claimed on the *EEBO-TCP* website[43], which also states that the digitisation process is carried out by typing, not by the less reliable method of OCR. These statements were in place at the time the play-texts were downloaded in 2009. However, it is not clear whether the percentage figure applies to every text or whether it is an overall average of accuracy of the *EEBO* digitised collection. Furthermore, according to Hope (2011), some of the files were originally digitised by OCR but are gradually being replaced with keyed versions. Hope's explanation would account for some of the problems I encountered with the digitised *EEBO* texts. I report these below in order to make clear my own editorial contribution to the *NDC* play-texts, and to alert other scholars wishing to use research methods which requires a high level of textual accuracy that would be undermined by missing words and typeset characters.

I noted a number of textual anomalies which, on investigation in the printed manuscript facsimiles, proved to correspond to missing typeset characters. These occurred in the form of spellings that did not look like EModE variants (e.g. "mistis" instead of "mistris"), or two consecutive typeset space characters (between which a word in the printed manuscript had been missed out in the digitised text). For example, in the digitised text of the line of print in Figure 4 on the next page (from Heywood's

[43]See http://www.textcreationpartnership.org/why-keying/ (last accessed 17.07.12).

play *Edward IV Part I*), the word "thunderclap" was missing in the digitised text, the only orthographic clue to this being two spaces between the words on either side:

Figure 4. Excerpt of *EEBO* facsimile printed manuscript from *Edward IV Part I*[44]

The missing word in this and other cases can probably be explained by unclear printing in the facsimile manuscripts, though in the text in Figure 4 "thunderclap" seems not appreciably less clear than the rest (to my eye, as a typist). This, together with the fact that the missing words in such cases have not been marked with leader dots in square brackets, suggests that the gaps are the result of OCR transcription. I located gaps by searching for double spaces in the play-text files and filling these in where possible from the printed manuscript facsimiles. Where I could not complete the gaps, I inserted a missing-text marker in a comment tag. The mis-spellings could not be searched automatically, but could be addressed by the spelling regularisation process (discussed in the next section).

The Early Modern printing convention of bracketing off the end of a line of dialogue longer than a standard page width, and putting it in the white space immediately above or below that line, appeared to have been mis-transcribed in a small number of the play-texts I used. In these, the bracketed line end had simply been transcribed as part of the line to which it was adjacent, rather than inserted into the correct speech turn. This occurred several times in the digitised text of Webster's *The White Devil*, an example of which is shown in Figure 5 on the next page.

[44] Image 2 of 88, *The first and second partes of King Edward the Fourth* (1600). STC (2nd ed.) / 13342. See http://eebo.chadwyck.com (accessed 20.10.10).

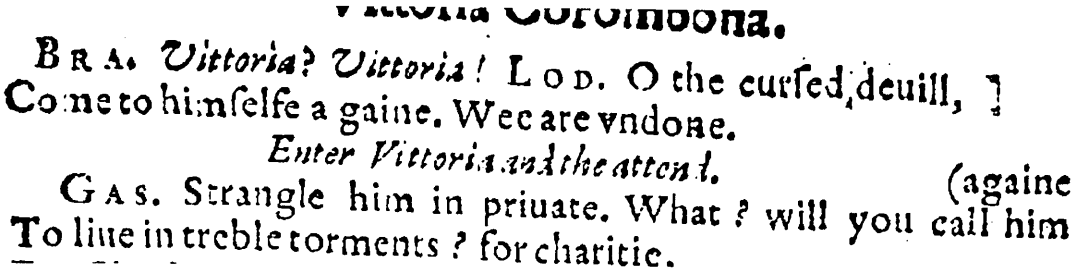

Vittoria Corombona.
BRA. Vittoria? Vittoria! LOD. O the curſed deuill,
Come to himſelfe a gaine. Wee are vndone.
Enter Vittoria and the attend.
GAS. Strangle him in priuate. What? will you call him (againe
To liue in treble torments? for charitie.

Figure 5. Excerpt of *EEBO* facsimile printed manuscript from *The White Devil*[45]

The bracketed text "(againe" in Figure 5 is actually the end of the line printed below, which is part of Gasparo's speech (labelled "GAS."). However, in the digitised text file it appears as a line on its own, immediately after the stage direction at the end of Lodovico's speech turn (labelled "LOD."), as shown in example (5) below.

(5)
```
LOD.
O the cursed deuill,
Come to himselfe a gaine. Wee are vndone.
Enter Vittoria and the attend.


(againe
GAS.
Strangle him in priuate. What? will you call him
To liue in treble torments? for charitie,
```

Webster, *The White Devil* (*NDC*)

Once I had identified this as a recurrent feature in one play-text, it was prudent to check all the others, in order to assess the extent of the problem. However, as indicated above, in most play-texts the dialogue affected by this EModE printing convention had been transcribed into the correct speech turn and word order.

Having downloaded the play-texts in 2009, I re-checked the text in example (5) on *EEBO* in July 2012, together with some unmarked missing words in Heywood's *Edward IV Part I* and Marlowe's *Tamburlaine Part I*, and found them unchanged. The potential impact of such anomalies would depend on the methods used in any

[45] Image 37 of 45, *The white diuel, or, The tragedy of Paulo Giordano Vrsini, Duke of Brachiano* (1612) See http://eebo.chadwyck.com (accessed 20.10.10).

particular study. My corpus linguistic methods require a high degree of accuracy of spelling and word order to maximise the retrieval of results, and I judged it necessary to check the badly-affected digitised texts on a line-by-line basis against the printed manuscript facsimiles. To minimise the need for this, I substituted a few play-texts which contained many transcription issues for others that were less affected, but equally appropriate for the corpus. However, as explained in 4.3, the process of compiling a suitable group of plays for the *NDC* was quite challenging, due to the need to balance dates, genres and other factors with the Shakespearean plays. Consequently, there was little room for substitution manoeuvres, and it was justifiable to invest time in improving the accuracy of the play-texts mentioned in this section, since they are of particular value to the contents of the corpus. Apart from maximising the potential for orthographic matching, and hence the retrieval of results, in principle it is desirable for the *NDC* play-texts to be as complete and faithful to the printed manuscripts as feasible. This is so they will better serve their purpose as a historical corpus, and ensure that my findings in this study and in the future can be deemed reliable.

I now turn to the regularisation of (authentic) EModE spelling variation in the *NDC* play-texts.

5.4 Addressing Early Modern English spelling variation in the play-texts

I mentioned in 4.2.2 that the play-texts in my study pre-date the completion of spelling standardisation in English[46]. As stated in 5.1, historical spelling variation hampers the retrieval of results using techniques that work by automated orthographic matching.

[46] See further Crystal (2008:43-48, 58-63) and Nevalainen (2006:4) with regard to spelling in Shakespeare's plays; see also Crystal (2012); Görlach (1991:8-9) and Scragg (2011 [1974]) for more detailed discussions of the development and standardisation of English spelling.

However, specialist software tools which regularise (or "standardise", or "normalise") multiple variant spellings of a word have, in recent years, made great headway in overcoming the obstacle of spelling variation in corpus research with historical texts. This enables linguists to investigate style features in texts from the Early Modern period using similar methods to those applied to more modern texts (with standardised spelling).

I follow Culpeper and Kytö (2010:112-113) in applying the *VARD 2* (*VARiant Detector*) software. This programme has been developed at Lancaster University, originally by Dawn Archer and Paul Rayson (see Archer and Rayson 2004; Rayson et al. 2005; Rayson et al. 2007), and more recently by Alistair Baron (see Baron and Rayson 2008, 2009). *VARD 2* has been shown to improve the potential for retrieving data in a corpus of EModE medical texts (by Lehto et al. 2010), and in the "Visualizing English Print from 1470-1800" project at the University of Strathclyde (Richard Whitt, personal communication, 16.07.12). As summarised in Baron and Rayson (2009: e.g. 2, 4), earlier versions of *VARD* were designed specifically for use with EModE texts, but more recent versions have evolved to address spelling variation in PDE texts such as e-mails, blogs and learner corpora as well. While this broadens the potential usefulness of the tool, it also creates some hazards that the historical researcher needs to watch out for.

VARD 2 finds multiple variants of a word and replaces them with a sole variant. It operates according to a combination of algorithms based on phonetic matching, letter-replacement rules and comparisons with dictionaries. This is explained in detail by Rayson et al. (2007:4-6, 9-10) and Baron and Rayson (2009:4-9), and summarised by Lehto et al. (2010). Lehto et al. (2010:286) explain that the reliability of the methods which govern *VARD 2*'s decisions to replace variants

appropriately depends on the corpus it is working with. The user can vary a "threshold confidence measure" option to control the amount of evidence the programme requires in order to make a replacement. *VARD 2* also has a training function through which it "learns" from sample data to make the kind of replacements which are desirable in a particular corpus. Training is strongly recommended by Baron (personal communication, 13.10.10), because text-types vary widely and the programme needs to be tailored to the language of a particular corpus to make more accurate replacements. To help with training and monitoring the effects of *VARD 2*, the regularised replacement variant and the original variant can be scrutinised if the XML output option is selected (plain text files are also an output option). Example (6) shows how a variant of *will* is regularised in my data.

(6) `<normalised orig="wil" auto="true">will</normalised>`

The main versions of *VARD 2* which have been successively released are documented at http://www.comp.lancs.ac.uk/~barona/vard2/versions.php[47]. An important difference between earlier and more recent versions is the replacement of the dictionary element to which the programme compares variants. V.2.1 and previous versions rely on a list of known EModE spelling variants, whereas V.2.2 and subsequent versions use a modern dictionary based on the *BNC* and *SCOWL* (*Spell Checking Oriented Word Lists*[48]; see Baron and Rayson 2009:4). The modern dictionary orients the programme to **modernisation**, rather than **standardisation**, of language: a crucial difference in historical research. The ability of V.2.1 to identify potential variants for replacement is limited to those which have been input by a researcher working specifically with EModE texts. (The list of known variants used by

[47] (last accessed 10.08.12).
[48] See also http://wordlist.sourceforge.net/ (last accessed 10.08.12).

V.2.1 was compiled manually during the development of UCREL's semantic tagging programme for EModE; see Archer et al. 2003.) Baron's rationale for modernisation is that if a word spelling is to be standardised it is more objective to use the PDE form than to choose one EModE variant or another (personal communication, 13.10.10). This is reasonable, but it means the historical researcher needs to ensure that the programme will not modernise spelling beyond a point which is useful for the study. Too much modernisation risks obliterating language features which are of interest, as I illustrate below.

In my study, the objective is to maximise the chances for word forms to be matched when the *SDC* and the *NDC* are compared. As the language in the play-texts of the *SDC* is already modernised to some extent, the amount and type of modernisation in the *NDC* play-texts needs to be on a par with it. Having used V.2.1.5 to regularise the spelling in the *SDC* play-texts in my (2009) research (as mentioned in 5.1), I tested this version with the *NDC* play-texts, but found that it left many words spelled with *i* instead of *j* (for example *iudge*) and *u* instead of *v* (for example *haue*). These are regularised in the *SDC* texts (even in their unVARDed condition) and would not therefore have been matched if left as EModE spellings in the *NDC*. V.2.3 replaced many more of them, but it also made some undesirable replacements which V.2.1.5 did not, such as regularising the pronoun *thou* to *you*, and modernising the verb forms which agree with *thou*. Since there is a distinction between the use of *thou* and *you* in the Early Modern period (with implications for social relations), they are not listed as EModE variants of the same pronoun in V.2.1.5's EModE dictionary, so the programme does not replace them. V.2.3, in contrast, automatically updates archaic forms which have corresponding modern forms in its dictionaries.

Following tests with both versions, I opted to use V.2.3 because some modernisation was useful to increase parity between the spelling in the *SDC* and the *NDC*. I employed the training facility of V.2.3 to modernise variant spellings more selectively, so as to improve precision and recall in retrieving results with less cost to the historical style features in the texts. I carried out the regularisation of spelling variation in the *NDC* with V.2.3 in three stages. First, I experimented to find the optimum confidence threshold at which the programme would make desirable replacements (such as the modernisation of words in which *i* is now *j* and *u* is now *v*), whilst minimising those that were undesirable. Baron and Rayson (2009:15) indicate that the 70% threshold is an appropriate choice to balance precision and recall, but I found this left too many words unregularised which had been modernised in the *SDC*. The 50% confidence threshold gave optimum results.

The second stage was to train V.2.3 at the 50% confidence level with samples of the *NDC* texts. *VARD 2* can be trained interactively, on a word-by-word basis (rather like spell-checkers in word-processing applications), but this is very time consuming and would not have been feasible in my study. A quicker option is to run the untrained *VARD 2* programme over a sample of the texts to be regularised, insert manual tags to correct or add to the regularisation as necessary, then re-load the trained texts into the programme. *VARD 2* picks up information in the manual tags and takes it into account when new texts are regularised. I used this training method. The training tags have a similar format to those inserted by the programme automatically (shown in example (6) above), but with the encoding "false" instead of "true" (Baron, personal communication, 13.10.10). This is illustrated in example (7), which shows the encoding for training the programme to regularise *greene* to *green*:

(7) `<normalised orig="greene" auto="false">green`

The extent to which *VARD 2* can learn through training is limited, however, because of the weighting of the algorithms which operate with the dictionary element, further details of which are given by Baron and Rayson (2009). They demonstrate that training improves the recall capability of *VARD 2* from 45% to 65% (2009:9-22), and they experiment with training *VARD 2* on different amounts of texts. Baron and Rayson find that recall improves consistently with training samples of up to about 12,000 words (tokens), at which 60% recall is achieved, but that it declines thereafter, increasing by only a further 5% when up to 40,000 words of training text are used (2009:9-14). In their study of EModE medical texts, Lehto et al. (2010:287-288) use 36,000 words of sample data (which they train with the alternative interactive method). As little testing of the training facility has been carried out in other studies, after discussion with Alistair Baron I trained *VARD 2* on a sample of 20,000 words from the *NDC* (or about 2.5% of the corpus), and then on a further sample of 20,000 words (about 5% overall), to compare the amount and quality of the replacements when the programme was then run on the remainder of the corpus.

As over 40,000 replacements were made each time, I did not check each one and quantify the improvements; instead, I spot-checked the replacements made by the programme each time and formed a judgement from the quality of these in terms of over-modernisation. There was a reduction in the number of errors made by the programme after the first training sample was re-loaded and the rest of the corpus processed, but no further improvement after the second training sample was re-loaded. I did not measure the improvement on a statistical basis, as my tests were not as detailed as those of Baron and Rayson (2009). Their goal was to quantify the benefits of recent developments to the programme itself, whereas mine was to assess the amount of benefit which the time invested in training the programme had on my own

data. However, my findings support their arguments that some training greatly improves the accuracy of *VARD* V.2.3, but that its capacity to be customised is limited. Clearly it is not worth investing more time in training than will pay off in increased accuracy.

Once the whole of the *NDC* had been processed with trained *VARD* V.2.3, the third stage in the regularisation of spelling was the manual correction of some inappropriate replacements which the training had not overcome. These were identified by checking samples line by line, noting frequently-occurring problem cases and amending them in the whole corpus (using *Notepad++*). It was notable that the regularisation of archaic forms (such as the replacement of *didst* with *did*) was not very well addressed by the training; the modernisation of these had to be reversed as *didst* is present in the *SDC*. Frequently-occurring homophones which are homographs in the *NDC* texts but not in the *SDC* texts were regularised at this stage. For example, *deere* was regularised to *dear* or *deer*, and *bee* was regularised to the verb form *be* in almost all cases, apart from a few instances where it occurs as a noun.

Non-standard punctuation in the *NDC* created complications which could only be addressed in a limited way, and manually. For example, *Ile* was regularised to *I'll* or *isle*, as appropriate for the context. In general, however, it is difficult to determine the standard form(s) to which punctuation should be regularised in this historical period. For example, the modernised *SDC* play-texts include apostrophes in s-genitives as standard, but the *NDC* texts from the late 16th and early 17th centuries do not (though there are some). Salmon (1999:46-47) argues that one of the compositors of Shakespeare's First Folio (dated 1623) uses apostrophes for the singular s-genitive. Nevalainen (2006:72) argues that the apostrophe does not actually occur in singular s-genitives until the second half of the 17th century, but Görlach states that:

> As regards spelling, the use of the apostrophe *(boy's)* was optional from 1500, frequent in the seventeenth century and fully established by 1690-1700 – the plural marking (*boys'*) was to follow only in the eighteenth century. (1991:82)

The arguments of the above historical linguists regarding the introduction of apostrophes to s-genitives vary somewhat, but indicate that they were rare, and probably used by compositors with a personal preference for them (perhaps because they were innovative and modern). As I argued in 4.2.2, compositors' preferences in the editing and printing processes are not necessarily those of the dramatists. This would be impossible to verify without consulting the original manuscripts of the plays: an unfeasibly large task in the present study, and also impossible in the case of plays for which no manuscripts survive. Analyses of authorial styles based on results containing contractions and punctuated language in the play-texts cannot therefore be reliably carried out in my study, so I do not follow up results with punctuation which occur as statistically significant (e.g. in 8.2). These are very few in my data, however, so it is only a minor problem.

In total, 43,193 variant forms in the *NDC* were regularised, amounting to just over 5% of the words in the corpus. 42,376 of the replacements were carried out by *VARD 2* after training, and 817 were performed manually (these figures exclude some "mistakes" made by *VARD 2* which were then reversed manually). The spelling regularisation process affects the word count of each corpus slightly, since some EModE forms are split (e.g. *shalbe* into *shall be*) and others combined (e.g. *an other* into *another*). *VARD 2* did not regularise compound forms such as *to morrow*, *your self* and *it self* to single forms, and I did not adjust them manually because in some cases it seemed that the meaning would be altered. I erred on the side of editorial caution and left them in their original compound forms, but this means that they cannot be considered reliable results. In this thesis, word counts for the play-texts are

based on the VARDed texts. VARDing the *SDC* reduced the overall count by 48 words (using V.2.1.5 for the 2009 study, and having excluded *Pericles*; see 4.2), whereas it increased the *NDC* by 2,938 words (using V.2.3 in the present study). The much lower number of replacements in the *SDC* mainly reflects the existing modernisation of the language in the texts before they were VARDed, although it is possible that more would be made if the *SDC* were VARDed with V.2.3, with its orientation to modernisation. Word counts of each play in the corpora, before and after VARDing, are given in Appendix IV.

The processing time of *VARD* V.2.3 is much improved over V.2.1.5, which took several days and nights to regularise the spelling in about 800,000 words of Shakespeare's plays (for my 2009 project). V.2.3 took about 15 minutes to process the same amount of data from the *NDC* for the present study.

The regularisation of spelling completed the preparation of the *NDC* play-texts for analysis with corpus tools. The processes described in this chapter represent a substantial investment of time and effort, so in the final section I briefly evaluate the main outcomes of my experiences using some relatively new software for annotation and for spelling regularisation, as well as the recently-digitised *EEBO* texts.

5.5 Discussion and conclusions

The annotation of the *NDC* play-texts using XML tags (discussed in 5.2) means that the socio-pragmatic and contextual information which is essential for interpreting my quantitative results in chapters 6 to 8 is successfully preserved, whilst necessarily excluded from computations made by the corpus tools. The amount and type of annotation is tailored to the needs of the present study, but with an eye to future research directions such as the analysis of language styles of characters of different

gender. PHP usefully enabled the rapid annotation of most of the speaker-id tags, but could not have been applied without in-house expertise in creating the scripts. It is not an intuitive process, although it is potentially powerful, and the user needs a relatively high level of programming skills in order to carry out a single command using PHP (compared to other software typically used in the preparation of corpus texts, which I mention below).

Furthermore, because PHP scripts are run from the command line, it is not possible to see the effects of a PHP script in the text(s) until the command has been completely executed. It is therefore necessary to re-save a new version of the text file in case it has not worked in the manner intended. It offers the advantage of being able to annotate multiple texts at the same time, unlike Microsoft's *Notepad*, *Word* and *WordPad* programmes, and a full range of regular expressions can be applied with it (again unlike the Microsoft programmes, although *Word* does offer some basic wildcards and limited regular expression capabilities).

Notepad++, the text editor I used with PHP, is itself a potentially useful programme by which to carry out corpus annotation. It is more intuitive, having drop-down menus and dialogue box options on a graphical user interface and, like PHP, it offers the option of editing multiple documents simultaneously, with full regular expression capabilities for searching and replacing. However, it is limited to executing one command at a time, whereas multiple commands can be built into a PHP script and executed simultaneously. *Notepad++* offers another advantage over PHP (in my view), in that its commands are carried out with the corpus text files open in the programme. They can therefore be tested on one or two cases in a single text, to ensure the regular expression has the desired effect, then applied to the entire corpus with a few further mouse clicks. As my discussions of annotating speaker-id tags in 5.2 show,

it is difficult to make global replacements without including some inappropriate candidates, but also to capture all the cases that need replacing. I find this easier with the text viewable on-screen.

It is possible that the size or number of texts that can be opened simultaneously in *Notepad++* is finite, so it may not be suitable for editing very large corpora, but it had no problem processing the 43 *NDC* play-texts all at once. I also encountered no problems using it to search and view particularly interesting results during the course of my analyses, with all the plays in both corpora open at once (79 documents comprising about 2 million words, taking into account the tagged text as well as the dialogue). *Notepad++* has a basic concordance feature which counts and displays results, although it does not offer a link from the concordance lines into the body of the text. For the corpus researcher without the resources to acquire the necessary programming skills for writing PHP scripts, I would suggest *Notepad++* as a useful tool for corpus annotation.

The *EEBO-TCP* is without doubt producing a resource of enormously valuable electronically searchable historical English texts, and I could not have built the *NDC* without them. Downloading, checking and amending the digitised *EEBO* text files was still vastly more efficient in terms of time and accuracy than copy-typing or scanning manuscripts of play-texts and checking them all would have been. Moreover, although time-consuming, the need to check and correct some of the plays on a line-by-line basis did provide an opportunity to get to know them better, and to enlarge my understanding of Early Modern printing and historical spelling conventions. Since the *EEBO-TCP* is an ongoing project, the minority of play-texts with missing words and typeset characters such as those noted in 5.3 can be expected to be replaced with more accurate versions (according to Hope 2011).

The training of *VARD* V.2.3 explained in 5.4, combined with some manual regularisation, successfully increased the parity between the spelling in the *NDC* and the *SDC*. The benefits of spelling regularisation of the *NDC* texts in this study cannot be quantified, but have been verified by checking samples of data (supported by the evidence of increased prospects for retrieving corpus results in other studies). *VARD 2* is a great boon to the researcher working with corpora of historical texts, affording the possibility of smoothing out a considerable amount of spelling variation which would otherwise undermine the retrieval of results extracted by automatic orthographic matching processes. The training feature in the newer version (V.2.3) is sophisticated, enabling corpus-specific tailoring to be built in to its search-and-replace processes. However, the replacement of EModE dictionaries in previous versions with modern dictionaries probably causes more training to be required with EModE texts than was necessary with the older versions such as V.2.1.5.

The modern dictionaries in more recent versions which extend the application of *VARD 2* beyond historical texts cause some disadvantage for historical researchers. Training *VARD 2* for a historical study now has to overcome the eradication of archaic forms which were standard in EModE (such as *thou* and verb forms which agree with it), in addition to tailoring the programme to particular EModE text-types. Regularisation with previous versions did, of course, involve standardising older forms to modern forms in many cases, but the tendency for V.2.3 to replace standard archaic forms with modern equivalents seems to have been caused by the introduction of modern dictionaries and the weight they carry in the programme's decision-making processes. I found this tricky to overcome through training and by adjusting the confidence threshold, although overall the programme made mostly desirable replacements through modernisation. A researcher wanting to preserve historical

language forms as much as possible for analysis may prefer to use V.2.1.5, with its limited but specialised EModE dictionary and its slow processing time, although it replaces fewer variants. Where modernisation is less consequential (or even desirable, as in my present study), V.2.3 is likely to be the better choice, subject to training and careful checking to ensure that the kinds of language features one wishes to study do not get erased.

The efforts to address transcription inaccuracies and spelling variation which I have discussed in this chapter may well be less important in studies using different methods. For example, if the aim is to search for particular words in the corpus which are known in advance, including the spelling variants, sufficient data might be obtained without investing resources in some of the processes I have applied. In a study like mine, however, these matters are potentially critical, since the analyses are determined by the results: the research is, as stated in 2.3, data-driven. For the data to indicate reliably what is in the corpus, and thereby provide the best evidence on which to base the analyses, it is therefore crucial that:

(i) the dialogic text is scrupulously excluded through annotation;

(ii) the texts are as faithful to the printed manuscripts as possible; and

(iii) spelling variation is modified in a way that maximises retrieval prospects whilst minimising the eradication of other historical language features.

Having explained my efforts to achieve the above, I now present my results and analyses in the next three chapters.

CHAPTER 6. INVESTIGATING SIMILARITIES AND DIFFERENCES BETWEEN SHAKESPEARE'S PLAYS AND OTHER CONTEMPORANEOUS PLAYS USING HIGH-FREQUENCY WORDS, WORD CLUSTERS AND SEMANTIC DOMAINS

6.1 Introduction

This chapter is the first of three in which I analyse the contents of the *SDC* and the *NDC*. The three chapters all follow a similar structure, and much of the introductory comment in this chapter also applies to chapters 7 and 8. In this chapter, I begin by investigating the evidence for similarities and differences in the language of Shakespeare's plays and other contemporaneous plays when my two corpora are examined independently of one another, using simple frequency. First I examine single words (in 6.2), then 3-word clusters (in 6.3), and finally semantic domains (in 6.4). Following the frequency comparisons discussed in this chapter, I move on to focus specifically on similarities, with the locked results in chapter 7, which provide a background against which a closer examination of differences can then be set in chapter 8, using key results.

My research aims and questions, set out in 1.2 and 1.3 respectively, encompass matters of style and methodology (including corpus-building), and these are all considered in the discussions of my analyses in this and the next two chapters. To address research question 1, I consider the implications for authorial style, in terms of preferences for language features which Shakespeare and other contemporaneous dramatists shared or did not share, based on evidence from my data. I also consider the extent to which these appear to be features which characterise the register of EModE drama, i.e. those which are used for their function rather than for aesthetic reasons (bearing in mind the distinction between style and register discussed in 2.2). There is more to say about some results than others, as is inevitably the case with frequency-based results of the kind in this study (as indicated in 2.7), and I carry out some longer

analyses of particularly interesting results as informal case studies. I bring in other relevant studies where appropriate and useful, from linguistic and literary research.

I comment on the quality, reliability and usefulness of the output produced by the different corpus methods, which contributes to answering research question 2. My discussions also take in any problematic results which arise from issues to do with compatibility of the corpora or the nature of the play-texts in them. This relates to research question 3, and is an important quality control exercise, bearing in mind the considerable challenges involved in compiling and preparing a parallel reference corpus for Shakespeare's plays (discussed in chapters 4 and 5). It is essential:

(i) to ensure that the results used for stylistic analysis of language style are not biased by choices made about what to include in the corpora; and

(ii) to establish whether any results arise simply because of problems with non-standardised spellings which reduce the potential for orthographic matching, when computations are made by the corpus linguistic software tools.

This is in order to avoid the pitfall of following up less useful or potentially unreliable results in more detailed qualitative analyses. Following the presentation of the results and analyses, I end each chapter with a brief further discussion of the main findings, and some conclusions about what the analyses have contributed to each of my main research aims (in this chapter, in 6.5).

As explained in chapters 2 and 3, the single word and 3-word cluster results are extracted using *WordSmith*, and the semantic domain results are generated with the USAS tool in *Wmatrix*. These are presented in tables which, due to limited space, include only the top 20 results for word and word cluster results and the top 10 results for semantic domains (domain results pattern into fewer categories, as observed by Rayson 2008, but they contain more words which need to be considered and

discussed). In presenting limited numbers of the most statistically significant results from my data, I follow other scholars, e.g. Culpeper and Kytö (2010:116-117). I also follow Culpeper and Kytö (2010) in adopting the principle that, if multiple results rank equally in the 10th or 20th position, all of them are shown (so the top 20 could in fact include 22 results, for example). This is preferable to dropping equally quantitatively significant results simply because of the order in which they are displayed by the software. However, no results actually tie for 10th or 20th position in my datasets.

In addition to raw frequencies ("RF"), I also give frequencies which are normalised per 10,000 words ("NF") in this chapter, because the sizes of the corpora are close but not identical (as detailed in 4.4). Rayson (2008:531) points out that results do not necessarily have equal statistical frequency in two corpora, just because they rank at a similar level on frequency lists: hence the need for normalised frequencies. Normalised frequencies of words occurring in both corpora could be tested for statistical significance, but *WordSmith* and *Wmatrix* do this automatically, when the locked and key results are generated in the next two chapters, so I do not do not do so here.

As stated in 3.5, I discuss the distribution of results where it is particularly unusual or noteworthy, but to conserve space I do not provide details of distribution of every result. As indicated in 1.4, investigating high-frequency language features in the two corpora in their entireties provides a broad, overall picture of the language in dramatic dialogue by Shakespeare and the other contemporaneous dramatists, obtained through the application of systematic methodology. This provides a robust starting point from which to begin discussing similarities and differences between Shakespeare and other playwrights of the period, but by no means an end point. Accordingly, I

mention possibilities for extending the analyses in my discussions in chapters 6 to 8, and in more detail in my conclusions in 9.4.

6.2 High-frequency words

Table 12 compares the top 20 most frequently-occurring words in Shakespeare's plays (from the *SDC*), and in the other contemporaneous plays (from the *NDC*).

Table 12. Top 20 rank-ordered words in Shakespeare's plays and other contemporaneous plays (minimum frequency=200)

Rank	Shakespeare's plays	RF	NF	Other contemporaneous plays	RF	NF
1	THE	26,316	330.2	THE	24,511	308
2	AND	23,630	296.5	AND	23,728	298
3	I	20,179	253.2	I	21,099	265
4	TO	18,164	227.9	TO	19,781	248
5	OF	15,696	196.9	OF	15,401	193
6	A	13,642	171.2	A	14,517	182
7	YOU	13,506	169.4	YOU	13,546	170
8	MY	12,042	151.1	MY	12,851	161
9	THAT	10,555	132.4	IN	10,919	137
10	IN	10,417	130.7	THAT	10,445	131
11	IS	9,122	114.4	IT	9,029	113
12	NOT	8,312	104.3	IS	8,790	110
13	IT	8,145	102.2	NOT	8,014	101
14	ME	7,592	95.3	FOR	7,923	99
15	FOR	7,456	93.5	ME	7,578	95
16	WITH	7,024	88.1	BE	7,467	94
17	BE	6,731	84.4	YOUR	7,417	93
18	YOUR	6,561	82.3	WITH	6,967	87
19	THIS	6,458	81.0	THIS	6,771	85
20	HIS	6,422	80.6	BUT	6,621	83

As Table 12 shows, apart from those ranking in 20th position, the most frequent words in each corpus are identical, and in very similar rank order of frequency. In terms of the most frequently-occurring words which differ between the corpora, HIS, which ranks 20th in the *SDC*, actually ranks 22nd in the *NDC* with a normalised frequency of 73.6 (not shown in Table 12). BUT, which ranks 20th in the *NDC*, actually ranks 22nd in the *SDC* with a normalised frequency of 75.4 (again, not shown). Effectively,

therefore, the most frequent single words in plays by Shakespeare and by the other contemporaneous playwrights are the same, and they are all function words.

Similar function words occur in similar rank order with similar normalised frequencies across both corpora when broken down into genres (comedy, history and tragedy), and by gender of speaking characters. In both corpora, the top-ranking word used by female characters is I, which ranks in third place for male characters. In both corpora, men use the word THE most frequently. The high frequency of first and second person pronouns (I, MY, YOU and YOUR in both corpora, and ME in the *NDC*) are worth noting as markers of an "[i]nteractive" style, according to the "textual dimensions" proposed by Biber (1988:56-58), and typical of conversational language. Biber (1988:21) notes that first person pronouns have been argued as denoting "involvement" between speaker and addressee, e.g. in Chafe's (1982) study, and it is not hard to see that this would be a useful means of constructing dramatic dialogue so that it establishes relationships between characters. The function of dramatic dialogue is essentially to involve characters with one another, and with a theme or plotline, and to involve the audience with the play at a higher discourse level (in Short's 1996 model, illustrated in Fig. 1, 2.2). The high frequency of NOT in both corpora fits with the arguments of Biber (1988:245) and Tottie (1991:16-19) that negation is more typical of spoken than written language. I discuss its function in the play-texts further in the next section.

The similarity of words which occur most frequently in both corpora is reassuring from a compatibility point of view, though a high proportion of function words such as pronouns would not be unexpected in any corpus of dialogue or other interactional speech. As to whether they are of further analytical interest, Rayson (2008:531) argues that function words are inevitably the most frequently-occurring in

corpora, and "generally of no further interest to anyone trying to differentiate the content of two corpora." This may well be the case in a study of different registers. However, in principle, function words in corpora of texts from the same register, but by different authors, can be of value, as is demonstrated in computational stylistic studies. Craig (1999:210-211) finds that function words and other "common" words "offer a remarkably sensitive index of differences of style" in an investigation of diachronic change in Ben Jonson's writing style. Petersen (2010:156-160) also uses function words in her authorial attribution study, and Burrows (1987) demonstrates that patterns of different function words contribute to character construction in Jane Austen's novels. For example, in comparing the dialogue of a male character, Henry Tilney, and a female character, Isabella Thorpe, in *Northanger Abbey*, Burrows (1987:3) finds that "he uses the word 'the' a third more than she and 'of' almost twice as often as she: she, in turn, uses 'not' more than half as often again as he, and more than doubles his use of 'I'." The words *the*, *of*, *not* and *I* are also among those occurring frequently in my EModE play data, as shown in Table 12, and their relative densities would be worth comparing in studies of individual characters or groups of characters such as males and females. These are beyond the scope of the present study, apart from the brief gender and genre breakdowns mentioned above. Moreover, function words are of potential sociolinguistic interest, too. Social psychologist James Pennebaker's (2011) study of PDE reveals links between pronoun (and preposition) usage, personality type and social and emotional behaviour[49].

However, in practice in this study, the virtually identical lists of function words in Table 12 are of limited use in getting to the heart of similarities and differences

---

[49] More than three-quarters of the top 20 words in Pennebaker's (2011) PDE study are also among the top 20 in each of my EModE corpora. Although *the*, *of* and *I* are among the top 20 most frequently-used words in Pennebaker's study, *not* is absent. This is probably attributable to the fact that Pennebaker's data is based on non-interactional as well as interactional language.

between the styles of Shakespeare's language and that of his contemporaries, when the corpora are considered in their entireties. Therefore, to broaden the picture, I applied a "stop list" (i.e. a list of words to be excluded from computation) to the "Wordlist" tool in *WordSmith*, so as to ignore the function words and compare only the most frequently-occurring content words in both corpora. The exact list of words which should be on a stop list is debatable, and depends on the text-types and the researcher's particular requirements. For example, if a decision is made to exclude words which function as auxiliary verbs, what is to be done about those which also function as main verbs and/or nouns, e.g. *will*? The orthographic matching processes of the corpus linguistic software will not distinguish between them, and will exclude them all.

As the basis of my stop list, I used one provided by Andrew Wilson (Lancaster University), which has been applied with the *Multilingual Corpus Toolkit* programme (Piao et al. 2002) and which comprises PDE pronouns, prepositions and auxiliary verbs. Many of these PDE forms are present in the EModE texts in my corpora, and I also enlarged the list to include historical forms (e.g. *thou* and *hath*), based on those in Görlach (1991:85-88) and Nevalainen (2006:77, 85). Wilson's stop list does not include *shall* or *will*. I added *shall*, since it is a frequently-occurring modal verb in my corpora, but I did not add *will* since it also functions as a noun. In principle I did not want to exclude any function words which occur in the same forms in content word classes, so I also did not add the EModE auxiliary verb forms *wilt*, *art* and *dare* to the stop list (the *OED* confirms that they had noun functions by the period from which my data comes). The stop list does not contain all the elements of pronouns present in my corpora. As noted in 5.4, some reflexive pronouns in the *NDC* are in the form of compounds (e.g. *your self*); *your* is on the stop list but *self* is not, since it also functions as a noun (see Görlach 1991:86 for more on compound pronouns in EModE).

Contracted verb phrases such as *tis* and *I'll* are not excluded, although they contain parts of pronouns which are on the stop list (*it* and *I*). The stop list contains a total of 158 words, and is provided in Appendix III. It is worth noting that if the spelling in the corpora had not been regularised, additional variant forms would also need to have been added (e.g. *haue* as well as *have*).

The stop list was fairly successful in isolating the content words, and the top 20 most frequent in each corpus are shown in Table 13.

Table 13. Top 20 rank-ordered content words in Shakespeare's plays and other contemporaneous plays (minimum frequency=200)

Rank	Shakespeare's plays	RF	NF	Other contemporaneous plays	RF	NF
1	WILL	4,922	61.8	WILL	5,336	67.0
2	ALL	3,587	45.0	ALL	4,364	54.8
3	GOOD	2,838	35.6	NOW	3,596	45.1
4	NOW	2,739	34.4	SIR	2,833	35.6
5	LORD	2,667	33.5	COME	2,484	31.2
6	O	2,579	32.4	GOOD	2,414	30.3
7	COME	2,497	31.3	LORD	2,318	29.1
8	SIR	2,474	31.0	I'LL	2,289	28.7
9	WELL	2,216	27.8	SEE	2,281	28.6
10	LET	2,085	26.2	LOVE	2,160	27.1
11	LOVE	1,920	24.1	WELL	2,064	25.9
12	MAN	1,806	22.7	LET	1,969	24.7
13	I'LL	1,735	21.8	MAN	1,898	23.8
14	KNOW	1,655	20.8	KNOW	1,643	20.6
15	SAY	1,638	20.6	CAN	1,615	20.3
16	SEE	1,416	17.8	O	2,396	30.1
17	TIS	1,386	17.4	KING	1,505	18.9
18	GIVE	1,316	16.5	TAKE	1,364	17.1
19	KING	1,314	16.5	SAY	1,323	16.6
20	TOO	1,229	15.4	TOO	1,168	14.7

Over three quarters of the most frequent content words in the two corpora are the same, as Table 13 shows. The results suggest a slightly more verbal than nominal style overall; 10 of the results from each corpus have verb functions, whereas nouns are fewer (LOVE, ranked 10th in the *NDC*, is sometimes a noun and sometimes a verb). A relatively verbal style fits with Biber's (1988:105) "interactive or involved" dimension

of language, and this would clearly be effective in connecting characters to one another in a play through their dialogue. It can therefore be considered characteristic of the register of drama.

The concordance data reveals that some of the most frequent content words in Table 13 are useful for involving the audience with what is going on in the play (at the higher discourse level). For example, the verbs KNOW and SEE often convey characters' attitudes of belief or certainty at the on-stage level, expressed on the basis of knowledge or visual evidence, respectively. The normalised frequencies suggest that SEE is used rather less in Shakespeare's plays (this is clearer when the keywords are examined later in 8.2). Through the voicing of these verbs by the characters at the on-stage level, the audience also get to "see" and "know" what is necessary to understand the play. Example (8) shows one character voicing an inference made from the behaviour of another, based on what he "sees".

(8) Iachimo: But I <u>see</u> you have some religion in you, that you fear.

Shakespeare, *Cymbeline*, I:iv (*SDC*)

Example (10) shows a character making a pejorative assertion about another, by sharing her knowledge with the audience.

(9) Dame Purecraft: but I <u>know</u> him to be the capital Knave of the land

Jonson, *Bartholomew Fair*, V:ii (*NDC*)

SAY is another high-frequency verb which characters in both sets of plays use to report the dialogue of others, to elicit information (e.g. *What say you?*), and often to announce their own views or reiterate them more clearly (e.g. *I say*).

Whereas the high-frequency verbs noted above are used by Shakespeare and the other contemporaneous playwrights to help engage the audience with the play through the characters' dialogue, the verbs COME and LET often help move the action

of the play forwards. Characters use them to invite one another to go somewhere or do something, setting up opportunities for them to depart from the stage. Sometimes both verbs are used, e.g. in constructions such as *Come, let's away* or *Come, let us go*. The comings and goings of characters constitute an important practical aspect of stagecraft. This is discussed in more detail by Herman (1995:159-162), who points out that:

> A dramatist's floor management strategies have to attend to the boundaries of interactions. Relevant characters have to be brought on-stage and taken off it; thus, incoming and outgoing personae have to be either incorporated into the speech already in progress, or be disengaged from it. (1995:159)

Herman (1995:160) discusses a variety of language strategies which Shakespeare uses in arranging the entrance and exit of characters. My data indicates that verbal strategies with COME and LET are particularly frequent, and also popular with other dramatists of the same period.

Moving on from verbs, Table 13 shows that there are similarities in the frequency of GOOD and LORD, which often co-occur in vocatives in both corpora (e.g. *good my lord, my good lord*). SIR, another vocative, also occurs with similar frequency. This partly reflects the fact that both corpora contain similarly high proportions of male dialogue and male characters (see 4.4). These vocatives also reflect conventions of politeness, deference and hierarchical social relations in the Early Modern period, and the fact that there are a lot of high status characters in both corpora who merit such terms of reference or address.

Some of the words in Table 13 are clearly there because of what the plays are about. KING is highly frequent in both corpora, which can be attributed to the number of storylines in the history plays which feature male monarchs. The concordance data shows that MAN arises mainly because discussions surround predominantly male characters, but also because it is used as a generalisation for humans or people (including women) by both men and women in both corpora. The *OED* indicates that

this was a conventional generalisation before the 20th century, although the strength of it in the corpora may also be influenced by the fact that the plays were all authored by men (see 4.3.2.3).

WELL, in Table 13, is a versatile word which functions variously as a noun, adjective, adverb and as an interjection or pragmatic marker in both corpora. O is a frequent pragmatic marker in both corpora. Its functions and variant forms in EModE are discussed in considerable detail by Culpeper and Kytö (2010:214-218, 238-243, 260-283). Although O appears to be rather less frequent in the other contemporaneous plays, when the raw frequencies of the spelling variant *Oh* are also considered (18 in the *SDC* and 805 in the *NDC*), the normalised frequencies level out to 32.4 in the *SDC* and 30.1 in the *NDC*. This variance in spelling was not adjusted by the regularisation process (discussed in 5.4), although a further variant identified by Culpeper and Kytö, *ô*, was replaced with *O*. These decisions were made on the basis that Culpeper and Kytö (2010:275, 277) argue that *ô* is a variant of *O* (the circumflex being a feature used by some dramatists, notably Heywood and Jonson, but not others), whereas *Oh* and *O* differ in some respects in the contexts in which they are used. *O* more often precedes a vocative, and is less often followed by an exclamation mark than *Oh* (2010:278). *O* in my data is discussed a little further in 8.2.

The high frequency of LOVE in both corpora reflects its importance in EModE plays (the concept of love in Shakespearean drama is the focus of Archer et al.'s 2009 corpus study, as noted in 2.5.4). Although Table 13 shows that GIVE and TAKE occur as the 18th most frequent content words in the *SDC* and the *NDC*, respectively, this is not as interestingly contrastive as it appears, as both words occur with similar normalised frequencies in each corpus, just not quite sufficiently frequently to rank among the top 20.

In all, therefore, the most frequently-occurring content words are, like the function words (in Table 12), very similar in both type and frequency in the *SDC* and the *NDC*. While the function words showed that the language in both corpora bears similar hallmarks of interactional speech, focusing on the content words has additionally shown that Shakespeare employed some dialogic strategies that were also typical of his contemporaries in:

(i) constructing dialogue that communicates the story of the play (through verbs which voice background, motives, plot and action); and

(ii) conveying the deferential social relations between the kinds of characters who feature most notably in the plots (men of high social status).

I now examine what is revealed by recurrent combinations of words in both corpora, by examining the most frequently-occurring 3-word clusters.

6.3 High-frequency 3-word clusters

The top 20 most frequently-occurring 3-word clusters in Shakespeare's plays and the other contemporaneous plays are shown in Table 14 on the next page. Those which are common to both corpora are highlighted in bold text.

Table 14. Top 20 rank-ordered 3-word clusters in Shakespeare's plays and other contemporaneous plays (minimum frequency=50)

Rank	Shakespeare's plays	RF	NF	Other contemporaneous plays	RF	NF
1	**I PRAY YOU**	242	3.0	**I WILL NOT**	215	2.7
2	**I WILL NOT**	213	2.7	**IT IS A**	187	2.3
3	**I KNOW NOT**	159	2.0	**IT IS NOT**	163	2.0
4	**I DO NOT**	157	2.0	**I DO NOT**	155	1.9
5	**I AM A**	139	1.7	**AND I WILL**	154	1.9
6	**I AM NOT**	137	1.7	**I PRAY YOU**	141	1.8
7	MY GOOD LORD	131	1.6	**I AM A**	138	1.7
8	**AND I WILL**	128	1.6	**I KNOW NOT**	125	1.6
9	**I WOULD NOT**	126	1.6	IT IS THE	124	1.6
10	**THIS IS THE**	120	1.5	**I AM NOT**	117	1.5
11	THERE IS NO	118	1.5	**I WOULD NOT**	108	1.4
12	**IT IS A**	115	1.4	**THIS IS THE**	106	1.3
13	THE DUKE OF	111	1.4	TO BE A	101	1.3
14	THAT I HAVE	107	1.3	MY LORD OF	99	1.2
15	THAT I AM	102	1.3	IT MAY BE	97	1.2
16	**I HAVE A**	97	1.2	YOU ARE A	97	1.2
17	I WILL BE	97	1.2	ALL THE WORLD	95	1.2
18	MY LORD OF	97	1.2	AS I AM	95	1.2
19	**IT IS NOT**	96	1.2	**I HAVE A**	90	1.1
20	I THANK YOU	92	1.2	THAT I MAY	89	1.1

Table 14 shows that 12 of the top 20 most frequently-occurring 3-word clusters in each corpus, or 60%, are common to both of them. Amongst the top 10, 7 are common to both corpora (or 70%), these being I PRAY YOU, I WILL NOT, I KNOW NOT, I DO NOT, I AM A, I AM NOT and AND I WILL. These results can be compared with those in Culpeper's (2011:73) comparison of 3-word lexical bundles in Shakespeare's plays and other EModE drama (using the *KEMPE* corpus, discussed in 2.5.4).

Culpeper states that:

> Shakespeare's lexical bundles are distinguished by the fact that his top five most frequent bundles begin with the first person pronoun. Also, it is interesting to note that the most frequent three-word unit in Shakespeare's plays, 'I pray you', is something that is not characteristic of other Early Modern plays, other genres, or of course present day plays. (2011:73)

The top 5 results in my data in Table 14 confirm Culpeper's findings: all the Shakespearean top 5 clusters begin with *I*, compared to just two of the top 5 other

contemporaneous clusters. In Culpeper's other contemporaneous drama data, IT IS A and IT IS NOT occur in the top 5, as they do in my data (lending support to Scott and Tribble's 2006:64 argument that a "robust core" of results is likely to be obtained with different reference corpora, noted in 3.6). I PRAY YOU is among the top 10 other contemporaneous clusters in my data, in contrast to Culpeper's data, but it occurs only just over half as much in the other contemporaneous dialogue as in the Shakespearean dialogue, where it is the most frequent cluster. *I pray you* is a pragmatic marker of polite deference which is associated with requests in EModE (Culpeper and Archer 2008:74-16; see also Lutzky and Demmen, forthcoming). Since requests perform many useful functions in plays, it would not be surprising to identify an associated pragmatic marker such as *I pray you* as a register feature of drama in this period. Culpeper and Archer (2008:73) argue that "asserting and negotiating rights and obligations would seem to be a good way of producing dynamic dramatic dialogue, and of providing information to the audience about the social constraints that compel, vex or appease characters". However, my analysis of key clusters later on in 8.3 lends support to Culpeper's (2011) argument that *I pray you* is a Shakespearean authorial style feature, not a register feature.

Table 14 indicates that more high-frequency clusters associated with deference occur in the *SDC* than in the *NDC*. In the *SDC* data, there are terms of address and reference (MY GOOD LORD, MY LORD OF, THE DUKE OF) and a speech act of thanking (I THANK YOU). In the *NDC* data, apart from I PRAY YOU the only other apparently deferential cluster is the term of address/reference MY LORD OF. These are of course very general results; among the high-frequency content words (in the previous section) similar terms of deference occurred with comparable frequency in both corpora. Therefore, it could be that some of the deferential clusters occur more in

Shakespeare's plays because there are simply more higher-status characters in the corpus. Address and the use of vocatives in Shakespeare's plays has been the subject of other research (e.g. U. Busse 2002b; B. Busse 2006). Further exploration of the data from the *NDC* in similar ways (in a future study) would be useful in enlarging the picture, for what it would reveal about plays of the period and also for its implications for social behaviour at the time. Crystal (2008:223) argues that address forms (in Shakespeare's plays) constitute "a sensitive index of personal temperaments and relationships", and Mazzon (2009:14) (with regard to medieval drama) that "[t]he study of address is an important indicator of the kinds of politeness strategies used in a language community, and also an indirect source of insight into the way social relations are perceived and encoded within that community."

Overall, nearly two-thirds of the stock of most frequently-occurring 3-word clusters in the two corpora coincides, and is used with similar frequency. The cluster results in Table 14 combine some of the language features identified as notable among the high-frequency single word results in the previous section, notably first and second person pronouns, verbs which are instrumental in conveying characters' beliefs, wishes and general motivations, and the negative particle *not*. The *not* clusters highlight some apparently important functions of negation in EModE dramatic dialogue, which I illustrate next. There is not space here for a full discussion of negation in language, and its effects, but see for example Hidalgo Downing (2000:23-79) for a comprehensive summary, and see further Horn (2001), Jespersen (1917 [1962]), Tieken-Boon van Ostade et al. (1998), Tottie (1991) and van der Wouden (1997).

As stated in the previous section with regard to the high frequency of NOT in both corpora, negation has been shown to be a feature of spoken language (in PDE), by Biber (1988:245) and Tottie (1991:16-19). Tottie's (1991:31-44) findings account for

this through the higher frequency of "repetitions", "denials", "rejections", "questions" and "mental verbs" in spoken language[50]. There are examples of these kinds of negated phenomena in my data, in the contexts of the clusters I WILL NOT, I WOULD NOT, I DO NOT, IT IS NOT, I KNOW NOT and I AM NOT. Characters use them to talk about non-actions and non-states, i.e. things which they do not want, know, feel or like, and events, states or activities which are not taking place. B. Walker (2012:113, 207-209, 230-231) finds that the three main narrators in Julian Barnes' novel *Talking It Over*, from the late 20th century, also frequently use negation to describe the events of the story and how they feel about them. Negation therefore seems to be a speech-related phenomenon in the dialogue of literary texts.

This can be accounted for by Tottie's (1991:43) argument that negation "adds to the emotional character of what is said", through the contribution of mental verbs to speaker "involvement" (Biber 1988:105-108) and interactional language (Chafe 1982). Negation is a useful way of loading dialogue with emotion, which helps to fuel important aspects of drama (and prose fiction) such as conflict, excitement and tension. Some of the *not* clusters in my data (though by no means all) are associated with mental verbs, including I KNOW NOT, and I DO NOT, the latter often being followed by *know* or *think*. This is illustrated in example (10), by an excerpt from Webster's tragedy *The White Devil*, in which the male character Francisco offers the view that Vittoria Corombona is not capable of the murder of which she has been accused:

(10) Francisco: for my part
I do not think she hath a soul so black
To act a deed so bloody,

Webster, *The White Devil* (*NDC*)

---

[50] Tottie (1991:314) distinguishes between "denials" and "rejections" as follows: "Denial relates to propositions and is normally dependent on linguistic means for expression, while rejection is a pragmatic category, not dependent on language and not necessarily relating to propositions (although capable of being expressed in natural language)."

Other frequent clusters such as I WILL NOT and I WOULD NOT do not often collocate with mental verbs in my data, however. Therefore, although a finding of relatively frequent negation in EModE dramatic dialogue can be said to be typical of spoken language, there are other stylistic explanatory factors to be explored, over and above the construction of an involved or interactional style. Hidalgo Downing (2000:197), whose research into negation concerns prose fiction, argues that "negation is a natural foregrounding device typically used in discourse". This would also be a likely reason for its use in dramatic dialogue, in particular because it clearly has functions in characterisation. Using concepts from Culpeper's (2001:167-172) characterisation framework, the previous and next extracts from the corpora illustrate negation functioning as an "explicit cue" to characterisation, through "self-presentation" and "other-presentation". In example (10) above, Francisco's evaluation of Vittoria's potential as a murderer is an example of other-presentation (as indeed are examples (8) and (9) in the previous section, in which the characters voice what they see and know, respectively, about other characters). Example (11), which features I WILL NOT, involves self-presentation, as I explain below.

(11) Arragon: I will not choose what many men desire,
Because I will not jump with common spirits
And rank me with the barbarous multitude.

Shakespeare, *The Merchant of Venice*, II:ix (*SDC*)

In example (11), the male character Arragon wants to win the hand in marriage of the fair Portia, but to do so he needs to make the correct choice between several caskets, one of which holds a picture of her. He states what he intends **not** to do (select the gold casket, whose monetary value might attract the avaricious majority of "barbarous" admirers), rather than what he **does** intend to do (pick the less valuable silver casket, with the aim of demonstrating that he is not primarily interested in marrying for

money). Through this negation strategy, Arragon self-presents as a more worthy suitor, set apart from the rest. This is then exploited to comic effect, for Arragon opens the silver casket to find not Portia's picture, as he confidently anticipates, but "the portrait of a blinking idiot", as he puts it. His considerable indignance over this, and Portia's evident relief, creates amusement for the audience. The high frequency of negation in my results therefore seems not only to be related to the construction of speech-related language, but also to its use as a resource for dramatic effects: one which was exploited by Shakespeare and other dramatists of the period.

Apart from negation, a combination of verbal style and highly-frequent first and second person pronouns in the cluster data also contributes to a sense of involvement (Biber 1988:105-108). This was mentioned in the previous section with regard to high-frequency single words in the corpora, and argued as an essential quality of dramatic dialogue. Most of the results in Table 14 above are verb phrases, in both corpora, though a few noun phrases also feature (e.g. ALL THE WORLD, in the *NDC*, and the noun phrase fragment THE DUKE OF, in the *SDC*). Being, doing, having, knowing and willing (in the sense of volition) are essential verbal ways in which characters get across who is doing what in the plays, to one another on-stage and to the audience, at the higher discourse level.

A similar prevalence of verb phrases over noun and prepositional phrases (or fragments of these) is found by Culpeper and Kytö (2010:118-119) in their lexical bundle data from EModE drama. They argue that this gives rise to an "interactive" style which, they say, "is, of course, not surprising, given that a Play-text constructs a dynamic interaction for public entertainment"[51]. They also report few lexical bundles which form part of questions, as is the case in my data (there are none). Thirteen of the

[51] Culpeper and Kytö (2010:119) find that PDE drama is also characterised by many lexical bundles containing verb phrases.

clusters from one or both of my corpora in Table 14, i.e. about two-thirds of them, also occur among the most frequently-occurring lexical bundles from EModE play-texts listed in Culpeper and Kytö's (2010:116-117) data (IT IS A, AND I WILL, IT IS NOT, I HAVE A, I WILL NOT, I KNOW NOT, I WOULD NOT, I AM A, I AM NOT, I DO NOT, AS I AM, IT IS THE and I PRAY YOU). That my data coincides with the findings from another similar study of EModE plays, using a different corpus, is reassuring, since it indicates that the texts in my two corpora are fairly typical of the register and historical period. My data also supports Culpeper and Kytö's findings about EModE drama.

As discussed in 3.2.4, Culpeper and Kytö (2010:107-134) examine the functions of the lexical bundles in their data, using a comprehensive classification framework specially designed to accommodate (a) recurrent word combinations and (b) EModE, and I apply this to my very similar data in order to compare our results more closely. The functions of the top 20 3-word clusters from the *SDC* and the *NDC*, tabled above, are shown in Table 15 on the next page, with raw frequencies given in brackets after each cluster. (See Table 1, 3.2.4 for the complete list of functional categories.)

Table 15. Functions of top 20 3-word clusters in Shakespeare's plays and other contemporaneous plays

		Shakespeare's plays	**Other contemporaneous plays**
Interpersonal	Speech act-related	*Directive* I PRAY YOU (242) *Vocative* MY LORD OF (97) MY GOOD LORD (131) *Thanking* I THANK YOU (92)	*Directive* I PRAY YOU (141) *Vocative* MY LORD OF (99)
	Modalising	*Volition* I WILL NOT (213) I WOULD NOT (126) *Intention* AND I WILL (128)	*Volition* I WILL NOT (215) I WOULD NOT (108) *Ability* THAT I MAY (89) *Intention* AND I WILL (154) *Approximator/intensifier* ALL THE WORLD (95)
Textual	Organisational	*Informational elaboration* AS I AM (78)	*Informational elaboration* AS I AM (95)
Ideational	Topical	*People* THE DUKE OF (111) *States* I KNOW NOT (159) I DO NOT (157) I AM A (139) I AM NOT (137) THIS IS THE (120) IT IS A (115) IT IS NOT (96) THAT I HAVE (107) I HAVE A (97) THERE IS NO (118) THAT I AM (102)	*States* I KNOW NOT (125) I DO NOT (155) I AM A (138) I AM NOT (117) THIS IS THE (120) IT IS A (187) IT IS THE (124) IT IS NOT (163) TO BE A (101) I HAVE A (90) YOU ARE A (97)
Mixed		I WILL BE (97)	IT MAY BE (97)

Table 15 shows that the function of the largest proportion of clusters in both corpora is to describe states, in the Ideational: Topical category. The States clusters often feature the verbs *be*, *do* or *have*, and they are used both literally and metaphorically. States clusters are also very prevalent in Culpeper and Kytö's data. As they point out, through States clusters the characters "express a personal opinion about an aspect of the speaker's self or that of another person" (2010:133). These are, of course, essential aspects to the fabric of a play. Two examples from my data are given below, from a Shakespearean and a non-Shakespearean play-text, featuring different States clusters.

(12) Mark Antony: I have a ship Laden with gold

Shakespeare, *Antony and Cleopatra*, III:ix (*SDC*)

(13) Barabas: My name is Barabas; I am a Jew.

Marlowe, *The Jew of Malta*, V (*NDC*)

In example (12), Antony states that he has access to wealth as part of a persuasive strategy in his interaction with the other on-stage characters, which functions to move the conversation forward. However, in example (13), the audience already knows Barabas is Jewish; stating his race in addition to his name when introducing himself to another on-stage character emphasises his Jewishness as part of his social identity (which is central to the plot of the play). So, States clusters in my data have a potential function in character identity construction, as well as being part of the basic linguistic furniture of dramatic dialogue.

Table 15 shows that the Interpersonal categories are also quite well populated with results. This is again the case in Culpeper and Kytö's (2010:132) data, and they argue that the items "relate to the articulation of personal desires and negotiation of social relationships": another essential aspect of a play-text. The Speech-act-related

clusters in my data are particularly similar to their lexical bundles in the same category, in that they are often "utterance launchers". As explained in 3.2.1, utterance launchers are stored formulaic structures with a first person pronoun followed by a verb phrase (Biber et al. 1999:1073; see also Culpeper and Kytö 2010:140). Although an efficiency measure in unscripted language, the prevalence of utterance launchers in scripted dramatic dialogue helps to construct a conversational style. In my data, I PRAY YOU and I WILL NOT are examples which occur frequently. They are non-idiomatic and extremely flexible in the kinds of speech acts and discourse acts which they preface in dramatic dialogue.

Overall, there are similar numbers of Interpersonal clusters in the *SDC* and the *NDC*, though as Table 15 shows, there are more Speech-act-related clusters in Shakespeare's plays and more Modalising clusters in the other contemporaneous plays. The Interpersonal: Speech-act-related: Vocative clusters in Table 15 all contain the polite term of address *my lord*, and reflect the fact that many of the characters in both corpora are noblemen: a minor similarity which nevertheless helps confirm that the content of the *NDC* is similar to that of the *SDC*. It bears out the efforts (detailed in 4.3) to choose plays that would best ensure the likelihood of obtaining results revealing stylistic distinctions, not merely contrasts that arise from different topics and themes in Shakespeare's plays and the other contemporaneous plays.

The identical Interpersonal: Modalising: Volition clusters in both corpora, I WILL NOT and I WOULD NOT, are also worth noting, since they show more clearly the way characters' wishes and feelings are communicated to the audience (mentioned above as an essential function of dramatic dialogue). This is shown in examples (14) and (15), from the *SDC* and the *NDC*, respectively. In (14), near the end of the play, Juliet resists the Friar's urging to leave Romeo's body and get to a place of safety.

(14) Juliet: Go, get thee hence, for I will not away.

Shakespeare, *Romeo and Juliet*, V:iii (*SDC*)

In (15), early in the play, the brother of the widowed Duchess of Malfi reveals his opposition to the idea of her remarriage.

(15) Ferdinand: I would not have her marry again.

Webster, *The Duchess of Malfi*, I:ii (*NDC*)

The verb *will* can convey future intent as well as volition in EModE (Blake 2002:122-123; Nevalainen 2006:95-96), and in the case of the cluster I WILL BE there is not a clear proportion of instances used either way. Therefore, it is classified as having a Mixed function in Table 15.

Table 15 shows that Textual clusters are scarce in both my corpora. This again is also the case in Culpeper and Kytö's (2010:133) data, and they suggest that questions do not feature highly as there are other possible strategies in drama for divulging information (e.g. soliloquies). There may be a gender distinction here, however, because in my (2009:125-137) research I found that female characters had a strikingly high concentration of *wh*-question clusters among the Textual results in Shakespeare's plays. This is worth noting for future investigation, but I do not pursue it further in this study.

The analysis of the most frequently-occurring clusters in the Shakespearean and other contemporaneous plays has further illuminated what the most frequent single words in the corpus showed, in 6.2, particularly in the verbs which predominate. They are instrumental in communicating the play to the audience, and can be considered register features which help the text fulfil its purpose as a play. They are very versatile; the States clusters built around *be*, *do* and *have* in my data are adaptable to a wide range of referential tasks in conversation. Similarly, the Volition clusters built around

will and *would* can be used to introduce a huge range of beliefs, feelings and motivations.

I now look at the most frequently-occurring semantic domains into which the words in the corpora are grouped, to examine similarities and differences in the kinds of concepts which occur most often in the two corpora.

6.4 High-frequency semantic domains

Table 16 on the next page shows the most frequent semantic domains into which the words in the *SDC* and in the *NDC* are grouped by the *Wmatrix* USAS tool, when the corpora are examined independently. In fact, these top 10 domains account for more than half the total words in the *SDC* and in the *NDC*. The three most frequently-occurring words in each semantic domain are shown in italics below each category label, as examples. The 21 main categories assigned by USAS were shown in Table 2, in 3.3.1 (and see Appendix II for the full tagset). As stated in 3.3.1, I consider categories to be reliable if at least 50% of the member words are appropriately classified.

Table 16. Top 10 rank-ordered semantic domains in Shakespeare's plays and other contemporaneous plays (minimum frequency=200)

Rank	Shakespeare's plays	RF	NF	Other contemporaneous plays	RF	NF
1	Grammatical bin (Z5) e.g. *the, and, of*	203,384	2551.7	Grammatical bin (Z5) e.g. *the, and, to*	188,307	2363.9
2	Pronouns, etc. (Z8) e.g. *I, you, my*	142,558	1788.6	Pronouns, etc. (Z8) e.g. *I, you, my*	134,143	1684.0
3	Being (A3+) e.g. *is, be, was*	23,628	296.4	Being (A3+) e.g. *is, be, are*	21,890	274.8
4	Negative (Z6) e.g. *not, no, nor*	13,865	174.0	Unmatched (Z99) e.g. *sirra, 'hem, th'*	16,194	203.3
5	Power, organising (S7.1+) e.g. *lord, sir, king*	13,763	172.7	Power, organising (S7.1+) e.g. *sir, lord, king*	13,134	164.9
6	Personal names (Z1) e.g. *will, York, Warwick*	13,470	169.0	Negative (Z6) e.g. *not, no, nor*	12,694	159.4
7	Unmatched (Z99) e.g. *didst, csar, wouldst*	12,906	161.9	Time: general: future (T1.1.3) e.g. *will, shall, 'll*	11,871	149.0
8	Anatomy and physiology (B1) e.g. *heart, hand, blood*	11,887	149.1	General actions, making, etc. (A1.1.1) e.g. *do, make, made*	10,311	129.4
9	Time: general: future (T1.1.3) e.g. *will, shall, 'll*	11,154	139.9	Anatomy and physiology (B1) e.g. *heart, blood, eyes*	10,045	126.1
10	Speech acts (Q2.2) e.g. *say, tell, answer*	11,153	139.9	Moving, coming and going (M1) e.g. *come, go, leave*	9,846	123.6

Table 16 shows that when the words in the corpora are grouped according to semantic meaning, the profiles of those which occur with the highest frequency are very similar in both corpora. The top two categories of "Grammatical bin" and "Pronouns, etc." capture the function words which were identified as being among the most highly frequent in the corpora in 6.2 (Table 12), for example *the* and *and*, and *I* and *you*. It is not surprising that these two categories feature strongly in language comprised of interactional dialogue, nor is the strength of the "Being" domain, ranked third in Table 16. This contains forms of the verb *be*, which was also one of the top 20 most frequent words in both corpora. Characters talk to one another on stage about how things are, especially with regard to themselves, which effectively tells the story of the play to the audience (at the higher discourse level). In the same vein, they talk about their own

and others' actions, particularly using the verb *do*, which was noted among the high-frequency single words in 6.2 and also the word clusters in 6.3. *Do* is the most frequently-occurring word in the "General actions, making, etc." domain, ranking 8th in the *NDC*, and it would appear as the 11th category in the *SDC*, just missing the top 10, with a normalised frequency of 139.5.

The strength of the semantic domain "Time: general: future" in both corpora (9th in the *SDC* and 7th in the *NDC*, Table 16) is due to the classification of the highly frequent verbs *will* and *shall* (and the contraction *'ll*) into this category. These often express characters' volition and intention, however, as they talk about their wants and plans. Although these verbs are future-oriented, the concordance data suggest that the main semantic function of many of the instances would be better classified into one or more of the "Psychological actions, states and processes" USAS categories (tagged with the initial X) which encompass meanings of decisions, wants and plans.

Characters frequently reporting their own speech and that of others (using words such as *say*, *tell*, and *answer*) accounts for the strength of the semantic domain "Speech acts". This ranks 10th in the *SDC* (in Table 16), and would be 11th in the *NDC* with a slightly lower but not hugely dissimilar normalised frequency of 119.6. The domain of "Moving, coming and going", ranks 10th in the *NDC* and would be 12th in the *SDC* (with a similar normalised frequency of 126.4). It reflects the importance of characters discussing their own and others' movements, both on and off the stage, which has a practical function of carrying the plot forward, as well as helping to tell the story of the play. This was flagged earlier, in 6.2, through the high frequency of the verb *come* among the content words in both corpora.

The semantic domain "Personal names", which ranks 6th in the *SDC* with a normalised frequency of 169.0, does not appear in the top 10 domains of the *NDC*,

where it has a normalised frequency of only 116.2. However, this is because the spelling of personal names in the *NDC* is not regularised, and USAS puts variants of names it does not recognise into the "Unmatched" category, which as Table 16 shows is greater for the *NDC* than for the *SDC*. For example, there are 96 instances of the word *Sejanvs*, a variant spelling of the name *Sejanus* (a character in the play of that name by Ben Jonson), which are categorised as "Unmatched". As explained in 3.2.2, the texts of the *SDC* have been modernised to some extent, while those in the *NDC* have not, and this accounts for the contrast in the number of words which USAS could not place into semantic domains. The "Personal names" domain is not reliable, however, as some of the most frequently-occurring words included by USAS rarely or never function as such in the dialogue (e.g. the contracted forms *'hem* and *th'*, shown in Table 16, in the other contemporaneous plays). Table 16 shows that the most frequently-occurring word in this category in the *SDC* is *will*, but in the play-texts *will* functions usually as a verb, occasionally as a common noun, and only rarely as a proper noun. Amongst other problematic words in this category are the verb *say* and the preposition *within*.

The "Negative" semantic domain, ranking 4th in the *SDC* and 6th in the *NDC* (Table 16), contains the high-frequency word *not*, plus others such as *none*, *nothing* and *no*. Again, in general it confirms earlier indicators about negation in both sets of play-texts, in the high-frequency single-word result NOT (6.2) and the 3-word clusters containing *not* (6.3). Calderwood associates negation particularly with the character Hamlet, in Shakespeare's play of the same title, e.g. with regard to "ironic non-meanings, his disclosures of a not-self, his advertisements of his not-revenge" (1983:104). These claims are based on qualitative analysis, whereas the quantitative data in my study, whilst not disproving Calderwood's literary critical claims, puts them

into perspective by indicating that negation is fairly widespread in plays by Shakespeare and other dramatists of the period. Furthermore, as discussed in 6.3, negation is characteristic of spoken language, and is clearly useful in rhetorical strategies of denial and rejection which can also function as explicit cues to characterisation in the forms of self-and other-presentation.

The strength of the "Power, organising" domain in Table 16 is due to the frequency of words associated with noble or aristocratic characters in both corpora, e.g. *king*, *lord*, *master*, *mistress* and *sir*. Oddly, the word *madam* is classified into the domain of S2.1 "People: female" by the USAS tool, although its implications for deference and power relations between characters are similar to those of *sir*. In a more detailed investigation of specific semantic domains it would need reclassifying.

Finally among the domain results, the "Anatomy and physiology" category contains concepts of body parts or components such as *heart*, *hand, eyes* and *blood*, which are often used in metaphorical contexts (e.g. *of royal blood*). The heart is one of the sources of human body metaphors identified by Archer et al. (2009:153-154) in Shakespeare's plays, and my data shows that other EModE dramatists made use of similar metaphors. For instance, the heart is sometimes personified in order to express the attitudes, feelings and emotions of characters, as in examples (16) and (17), from the *SDC* and the *NDC*, respectively.

(16) Lady Macbeth: For my heart speaks they are welcome.

Shakespeare, *Macbeth,* III:iv (*SDC*)

(17) Sullen Shepherd: My heart has been her servant.

Fletcher, *The Faithful Shepherdess*, II:i (*NDC*)

Although the other high-frequency domains in Table 16 usefully confirm some of the similarities between the corpora that were identified through single words and clusters

in the previous two sections, the "Anatomy and physiology domain" is the only one in the top 10 which actually reveals something new. By grouping together words in the corpora according to the semantic concepts they concern, the USAS tool has highlighted the fact that some of the same kinds of metaphors are used by Shakespeare and by other contemporaneous dramatists. As argued in 3.3.3, the investigation of metaphor in EModE requires detailed and careful research into the surrounding context in order to infer accurately the meanings of metaphor in that period (based on Oncins-Martínez 2006 and Tissari 2010a), and also considerable manual classification of results. I do not pursue it further here, although a little more about metaphor emerges in the analyses of locked and key semantic domains in 7.4 and 8.4, respectively.

This completes my discussion of the most frequent semantic domains in the *SDC* and the *NDC*. I now give a brief summary and some conclusions about what the investigations in this chapter have contributed to my research aims.

6.5 Discussion and conclusions

6.5.1 Implications for Shakespeare's style and for the register of Early Modern English drama

The close similarity between the most frequent words, 3-word clusters and semantic domains in 6.2-6.4 have revealed little about the distinctiveness of Shakespeare's authorial style, but much more about the similarity between the language he and other contemporaneous dramatists use in constructing dramatic dialogue. The most frequently-occurring language features clearly have core functions associated with the register of EModE drama, specifically in:

(i) creating language which is conversational and interactional; and

(ii) communicating descriptions of people, events and the relationships between them in the form of a coherent and entertaining whole to an audience (or reader).

I summarise the main findings below.

- The most frequent words in both corpora are function words, particularly first and second person pronouns. These are typical features of "involvement" and "interaction", according to Biber (1988:56-58, 105), as is a relatively verbal (as opposed to a nominal) style, borne out by half the most frequent content words in Table 13 (in section 6.2). The 3-word cluster data from both corpora is characterised by formulaic "utterance launchers" (Biber et al. 1999:107; Culpeper and Kytö 2010:140), which also contribute to a conversational style (6.3).
- Shakespeare's characters and those in other contemporaneous plays frequently make statements about how things are, for example through clusters built around the verbs *be*, *do* and *have* (6.3), and in so doing communicate the events of the play to the audience. Notably, however, they also often state how things are not. This was shown in the high frequency of negated forms for all three types of language unit. This negation contributes not only to the spoken, interactive style (Biber 1988:105-108; Chafe 1982; Tottie 1991:43), but also helps charge the dialogue with emotion. Furthermore, as illustrated by some of the *not* cluster data (6.3), negation is instrumental in the self- and other-presentation of characters (Culpeper 2001:167-172).
- Characters in both corpora communicate intentions, wishes, desires, feelings and general motivations in similar ways, notably through the frequent volition cluster I WILL NOT (6.3).

- Shakespeare and other contemporaneous dramatists make frequent use of some similar language strategies in orchestrating characters' movements on and off the stage, noted in the high frequency of the verbs COME and LET (6.2).
- There is evidence of similarity in the most frequent content words which convey deferential social relations (KING, GOOD, LORD and SIR), but the 3-word clusters indicate that Shakespeare makes greater use of deferential language in comparison to his contemporaries (though this might be attributable to the social status of characters in the two sets of plays, and/or to particular kinds of speech acts such as requests).

Next, I assess what the high-frequency results have revealed about the compatibility of the new parallel corpus of other contemporaneous drama with the corpus of Shakespeare's First Folio, before ending the chapter with a reflection on the value of applying the method of analysing simple frequency counts in the study.

6.5.2 Corpus compatibility issues highlighted by the analyses of frequency

The close similarity between the most frequent words, 3-word clusters and semantically-linked groups of words in the *SDC* and the *NDC*, whilst not revealing a great deal about Shakespeare's language style in comparison to that of other dramatists, usefully indicates that there are no major compatibility issues between the play-texts. Those that do emerge are confined to certain categories of results only. I noted that variation in the spelling of the interjections *O* and *Oh* could not be fully addressed by the regularisation process (in 6.2), because there is evidence that the meanings were not entirely synonymous, but I also pointed out other claims that different dramatists preferred different spellings anyway (Culpeper 2010:275, 277). This would need to be

borne in mind if pragmatic markers in the corpora were studied in any detail, as would the compositors' preferences (noted as an issue in 5.4).

It was clear that greater levels of spelling standardisation in the relatively modern *SDC* texts enabled the *Wmatrix* USAS tool to match more personal names in Shakespeare's plays than in the other contemporaneous plays (in 6.4). However, this is not very material to my investigation of language style. Some of the high-frequency results which are vocatives indicate similarities in the topics and storylines in both sets of plays, such as KING, LORD and SIR (in 6.2), and the "Power" semantic domain (6.4). This is not surprising, but usefully helps to confirm that the efforts to balance the content in the parallel reference corpus to that in Shakespeare's First Folio as far as possible (explained in 4.3) have been generally successful.

6.5.3 Evaluating the method of frequency

The frequency counts of individual words, 3-word clusters and groups of words which are semantically related have, overall, provided a rudimentary comparison of the language in the *SDC* and the *NDC* at three levels. It was prudent to examine the corpora separately to start with, to note any issues in the most frequently-occurring language features in the results which might indicate compatibility problems. This was important to assess at the outset, because such problems can be masked by methods such as locking and keyness, which compare the corpora automatically to one another using statistical tests. Invisible compatibility issues would potentially lead to the misinterpretation of results. However, the close similarity of language features shown by the analyses in this chapter is reassuring. It is also not without interest in terms of examining the style of language in Shakespeare's plays compared to that of other dramatists, although the findings are essentially limited to similarities, and mainly

associated with the function of language in communicating a play. In other words, the results are mostly indicative of register features rather than authorial style features. Nevertheless, register features are created by repeated choices and preferences for certain ways of doing things with language. It is fair to say that Shakespeare and other popular and successful contemporaneous playwrights contributed to what now appear as register features of EModE drama, by reinforcing their individual style preferences. As argued in 2.2, style and register are not entirely clear-cut distinctions when examining the implications of particular language features.

In order to get nearer the heart of how and why the dramatic dialogue constructed by Shakespeare and his contemporaries is similar and different, I now go on to deeper explorations, with the benefit of statistical testing. In the next chapter, the locking method builds on the findings in this chapter, by taking the analysis of similarities much further.

CHAPTER 7. INVESTIGATING SIMILARITIES BETWEEN SHAKESPEARE'S PLAYS AND OTHER CONTEMPORANEOUS PLAYS USING LOCKWORDS, LOCKED WORD CLUSTERS AND LOCKED SEMANTIC DOMAINS

7.1 Introduction

In this chapter, I apply Baker's (2011) concept of statistical "locking", explained in 2.6, to identify the high-frequency words which occur statistically with the most similar frequency in the *SDC* and the *NDC*, in 7.2. I then extend this concept to 3-word clusters and semantic domains which lock across both corpora, in 7.3 and 7.4 respectively. The language features highlighted by the locked results reveal statistically-based shared preferences for words, clusters and semantic domains between Shakespeare and a group of his peers. In this chapter, I am examining the opposites of keywords, key word clusters and key semantic domains (statistically speaking; see Baker 2011:73 and my discussions in 2.6.2). In so doing I test out and evaluate the application of the locking method, in addition to furthering my investigation of the styles of Shakespeare and other contemporaneous dramatists. Following the analyses, I provide some further discussion and conclusions in 7.5.

7.2 Lockwords

Table 17 on the next page shows the top 20 lockwords, their raw frequencies in both corpora, and the p values (since the strength of the locking lies in the closeness of the p value to 1.0; see 2.6.2).

Table 17. Top 20 rank-ordered lockwords in Shakespeare's plays and other contemporaneous plays (minimum frequency=200; p=1.0)

Rank	Word	Shakespeare's plays	Other contemporaneous plays	p value
1	FELLOW	310	324	0.988
2	WAS	2,210	2,314	0.982
3	EVERY	497	519	0.974
4	YOUNG	414	432	0.968
5	ILL	259	270	0.966
6	DEATH	844	885	0.965
7	FAREWELL	355	373	0.955
8	I	20,179	21,099	0.940
9	GET	302	314	0.937
10	MIND	338	356	0.931
11	JOHN	268	278	0.919
12	WE'LL	310	327	0.919
13	THIS	6,458	6,771	0.908
14	IN	10,417	10,919	0.898
15	WORLD	597	630	0.882
16	PARDON	299	309	0.878
17	NAME	639	675	0.863
18	SET	443	469	0.860
19	POWER	324	334	0.848
20	MAJESTY	256	273	0.827

Table 17 shows that the words which are statistically "locked" across the two corpora are a very mixed bag, in comparison to the patterns of frequent verbs when the corpora were examined independently (in 6.2). As pointed out in 2.7, since there are no existing studies of lockwords in EModE drama, there was no basis from which to predict what they would be like (unlike key results, which have received much greater attention in corpus stylistics). In that sense, they are all unexpected, but some are more surprising than others.

One such surprising result is the most strongly locked word across both corpora, FELLOW, which Table 17 shows as having the most similar frequency (statistically). I would not have anticipated it among the results, because it does not feature very much in other studies which focus on address in EModE drama. With regard to Shakespeare's plays, B. Busse (2006:217, 228) categorises *fellow* as an

"epithet" in her study of vocatives, i.e. as a term which is distinct from conventional vocatives and which "describes a kind of quality already inherent in the semantics of the lexeme used", along with others such as *sirrah* and *friend* (2006:12). U. Busse (2002a) does not mention it, though in his (2002b) study he cites Barber's (1981:285) finding of it as a collocate of *thou* but not *you* in *Richard III*. Mazzon (2002:240, who also cites Barber's 1981:175 research), mentions *fellow* among terms of abuse which characters of higher social rank use to those of lower social rank, and Stein (2002:263) notes that "*fellow* is the usual designation for a member of the lower classes" (citing Replogle 1967:57). These scholars do not explore the use of *fellow* and its social implications any further in their analyses, however. Its position as the most strongly locked word in my data indicates that it warrants more attention in EModE drama than it has thus far attracted, so I investigate it a little further.

As with some of the most frequent words in 6.2, *fellow* is a term of address and reference used more about men than women (according to the *OED*, and supported by the evidence in my corpora). Table 18 shows that in my data the frequency of use by men and women (rather than about them) is similar for both corpora: male characters use *fellow* about twice as often as female characters.

Table 18. Frequency of use of *fellow* by male and female characters in both corpora

	RF	**NF**
Female characters in Shakespeare's plays	36	2.6
Female characters in other contemporaneous plays	36	2.3
Male characters in Shakespeare's plays	275	4.2
Male characters in other contemporaneous plays	281	4.4

Given the dominance of male characters and male dialogue in the Shakespearean and the other contemporaneous plays (see 4.4), it is not surprising to find a high frequency of male-oriented terms, used by characters of both sex in both corpora. However, unlike those which arose in the results in 6.2 (e.g. *lord* and *sir*), *fellow* is not a term

associated with deference, as is indicated by the above studies noting it as one which is derogatory. In my data, when it is used about women its meaning is that of a pair or counterpart. This is illustrated by an excerpt from Jonson's *Volpone*, in example (18). A character of high social rank, Lady Politic Would-be, asks one of her waiting-women about the whereabouts of the other.

(18) Lady Politic Would-be: Where's your fellow? call her.

Jonson, *Volpone*, III:iv (*NDC*)

Of the use of *fellow* in Shakespearean drama, Crystal and Crystal (2002:173) list eight distinctive senses, again mainly involving companionship or pairing, all of which have positive connotations except for "worthless individual, good-for-nothing". The collocates of *fellow* in my data are mostly pronouns, articles, and positively-connotated adjectives (e.g. *good*, *honest*, *sweet*, *pretty*, *brave*, *gallant* or *fine fellow*). A more detailed investigation of the use of *fellow* according to the social rank of speakers and addressees would be useful, but is outside the scope of this chapter. However, I would suggest that the notably similar high frequency of *fellow* in dialogue by Shakespeare and by other contemporaneous dramatists highlights it as a term of some importance in establishing character relationships and social relations. Although it may be the case that it is used towards lower-ranking characters, the evidence from the collocations in my data and from Crystal and Crystal (2002) indicates that it is not generalisable as a derogatory or abusive term – at least, not in the register of drama. A brief further investigation of its use outside of drama proved interesting, and is worth including for what it shows about the need for caution when interpreting (a) historical meanings and (b) meanings in a particular register.

According to the *OED*, the meanings of *fellow* during the late 16th and early 17th centuries were associated with comradeship, pairing, similarity and equivalence,

i.e. positive or neutral connotations, though it could be used in the pejorative sense of an accomplice (e.g. in a crime) and to indicate contempt. Onions (1982:80) and the *OED* state that an earlier sense of "polite condescension" conveyed in the use of *fellow* as an address term had been lost by the time Shakespeare was writing plays, but that it was a term appropriate only to a much lower-ranking addressee, such as a servant. However, apart from one attestation, all those given by the *OED* for the Early Modern period are from Shakespeare[52]; illustrating Benson's (2001:44) argument that "dictionaries create the distinctions they seek to record" (in Hope 2004:12). This is a case in point which supports Culpeper's (2011) call for a corpus-based, comparative dictionary of Shakespeare's plays (mentioned in 1.1), to widen the basis for interpretation of the language used in them.

The publicly-searchable version of Lancashire's (2012) *Lexicon of Early Modern English*[53] ("*LEME*") is much more fruitful, providing numerous attestations of *fellow* in diverse sources of the period (I used search dates of 1575-1640, i.e. slightly wider than those of the texts in the corpora, which returned 1,013 entries from 87 lexicons). These provide some further insight into the way *fellow* is used outside of the register of drama in EModE. The majority of the attestations in the *LEME* show *fellow* as having a pejorative meaning; for example, Henry Cockerham's (1623) *English Dictionary* records adjectives such as *lewd*, *base*, *crafty*, *cheating*, *rude* and *foolish* connotating with *fellow*. The *LEME* data indicates that in wider use the word *fellow* was more regularly used pejoratively than the proportion of pejorative examples in my EModE plays data would appear to reflect. This is interesting, and bears out the arguments of Mazzon (2002:239), who states that language use in [historical] literary texts does not necessarily represent what was habitual or conventional in wider usage,

---

[52] For a recent discussion of the citation of Shakespeare's work in the *OED*, see Goodland (2011).
[53] See http://leme.library.utoronto.ca/public/intro.cfm (accessed 23.03.12).

and Booth (2004 [1997]:20), who says that "Shakespeare's language and the language of Shakespeare's time were not the same".

The *LEME* displays a plot showing the distribution of the entries over the period searched, so I searched the entire *LEME* to see the diachronic distribution of *fellow* across all the lexicons it contains (dated between 1450 and 1702). This returned 1,444 entries from 176 lexicons, distributed as shown in Figure 6.

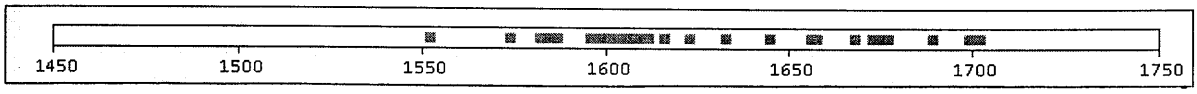


Figure 6. Diachronic distribution of entries for *fellow* in the *LEME*

Interestingly, Figure 6 shows a concentration in the entries for *fellow* just before and after 1600, coinciding roughly with the span of dates of the play-texts in both my corpora (1584-1626; see 4.3.2.1). The diachronic distribution plot needs to be interpreted cautiously, however, since the number of entries in lexicons of a particular period is by no means a direct indicator of frequency of use in the general population, or in any register other than the lexicons themselves. It is suggestive of the popularity of the word among dictionary writers and those who recorded and commented on language, though, which obviously has some correspondence with contemporaneous usage. It might reflect a period of change and/or innovation in the use of *fellow*, which dramatists were also in the vanguard of (from the strength of the word in both corpora, and also the fact that it is not overwhelmingly used pejoratively). However, there are issues of whether the size and number of lexicons is equally balanced across the whole period represented by the *LEME*; there may simply be more data from lexicons produced around 1600. A further question is whether the sources of the entries in the lexicons correspond to the dates of the lexicons themselves: in the restricted search of *fellow*, I noted a 1598 edition of the works of Geoffrey Chaucer among the sources, which of course represents language of a much earlier period. Notwithstanding these

cautionary comments, the *LEME* is a potentially useful resource, whose contents help enlarge the picture of the data from the corpora.

Among the other lockword results in Table 17 above, the second most strongly locked word WAS can be explained simply by the function of the past tense form of *to be* in narrating and describing what has taken place, which forms a large part of what most characters in both corpora are doing through their dialogue (to tell the story to the audience). Similarly, the fact that characters tend to talk about themselves is apparent in the lockwords I and WE'LL (the latter expressing future-oriented intent). JOHN seems to be just a popular name for male characters in plays of the period, but is worth noting as the single proper noun which occurs among the lockwords results. Proper nouns are also typical in keywords output (see 2.7, with reference to Scott 2000). PARDON is one of a range of polite ways to request "agreement or permission to do something" in Shakespearean drama, according to Crystal and Crystal (2002:314; 340-341). This is conventional in the context of hierarchical social relations around which all the plays revolve, and its presence as a lockword indicates that Shakespeare's characters use it in ways which are also typical of characters created by other dramatists of his era.

As cautiously anticipated in 2.7, some of the lockwords are there simply because of what the plays are about. They correspond with themes that are already known to permeate EModE drama, and which reflect contemporaneous social and political issues. For example, Heinemann (2003) and McRae (2003:105-121) discuss the ways political and monarchical power and change are played out in EModE plays, and those by Shakespeare and the dramatists in the *NDC* often centre around royal families, particularly in the history and tragedy genres. This accounts for the locking of the words MAJESTY and POWER across the corpora. The dispersion plots I

examined show a lighter distribution among the comedy plays. Power is a main driver of plots, particularly in history and tragedy plays of the Early Modern period, and it is something which characters discuss a lot. The popularity of revenge and retribution as plot drivers in EModE tragedy (discussed by Watson 2003:308-343, for example) accounts for the presence of DEATH as a theme which locks across both corpora (the dispersion plots confirming that it is again more heavily distributed in the history and tragedy plays). It is at least in part semantically linked to another lockword, ILL, which is a frequent descriptive term for a range of things that are bad, wrong or not as they should be, such as news, behaviour and omens. Whilst not revealing anything particularly new about EModE plays, these lockwords confirm statistically that the dialogue in the *SDC* and the *NDC* is similar in its orientation to certain common themes.

Several of the lockwords in Table 17 reflect popular ways for characters to talk about matters, and are worth commenting upon briefly. These are MIND, SET and WORLD. Crystal and Crystal (2002:282) and Onions (1982:140-141) indicate that *mind* has a function of expressing opinion and intention, e.g. in constructions such as *to my mind* and *I have a mind.* The concordance data shows that *mind* is often used to convey characters' motivations, wishes, beliefs and intentions. This figurative way of expressing such things is an interesting contrast to the more literal *will* and *shall*, noted in 6.4 as being of high frequency (in the semantic domain "Time: general: future"), and having similar functions of conveying essential information from playwright to audience. The concordance data shows that SET is most often used as a verb in both corpora, and is, as Onions (1982:194) states, "used in many connexions where 'place' or 'put' is now idiomatic". The *OED* and Crystal and Crystal (2002:393) broadly support what Onions says, and they additionally indicate that *set* can introduce a sense

of opposition, comparison, ranking or evaluation in the putting or placing of things. Some examples of the concordance data from both corpora are given below in Figures 7 and 8, from the *SDC* and the *NDC* respectively.

N	Concordance	File
1	t answer, nay; for indeed, who would set his wit to so foolish a bird? who woul	\scamids.txt
2	ur! I shall- Will you set your wit to a fool's?	\sctandc.txt
3	with the wild and wastefulocean. Now set the teeth and stretch the nostril wide,	1\shhenv.txt
4	r out the day in peace; but, ere sunset, Set armed discord 'twixt these perjured	r~1\shkj.txt
5	e the ides of March. Set him before me; let me see his face.	ar~1\stjc.txt
6	under his pillow, and halters in his pew; set ratsbane by his porridge; made him	~1\stlear.txt
7	lmostdamned in a fair wife; That never set a squadron in the field, Nor the divi	r~1\stoth.txt
8	hen plainly know my heart's dear love is set On the fair daughter of rich Capule	1\strandj.txt
9	d To bring manslaughter into form, and set quarrelling Upon the head of valour	1\sttimon.txt
10	these dire events. Set him breast-deep in earth, and famis	~1\sttitus.txt

Figure 7. Concordance extract for *set*: Shakespeare's plays

N	Concordance	File
1	an no further: what poor ass was it that set this in my way? now if my father sh	ncanhum.txt
2	in your belly, Nightingale,come, Sir, set it here, did not I bid you should get	2\ncbfair.txt
3	r? Thou set thy son to scoff and mock at me,	nctwoang.txt
4	ar me, lady? Why, if your knight have set you to beg shirts, Or to invite me	~2\ncvolp.txt
5	chariot wheels Restless,till I be safely set in shade Of some unhauntedplac	2\nhalcaz.txt
6	mmandme to give over holy day, And set wide open, what you would not see.	\nhdeath.txt
7	aning, there's a privy thief I know you set to pillage my affections, He durst	2\nhduch.txt
8	ng for Gods sake: truly if you do I shall set a knave betweenyou.	2\ntawkk.txt
9	ss, He was a pretty Poet too, and that set him forwards first; The Muses the	ntchange.txt
10	Drive if you can my house to Italy: I'll set the casementopen that the winds	~2\ntdido.txt

Figure 8. Concordance extract for *set*: other contemporaneous plays

The varied examples in the concordance data for SET show that it is a convenient way of expressing how things are arranged (literally or metaphorically), and of talking about purpose or design, all of which characters frequently do in dramatic dialogue. The shared high frequency of SET between both corpora, causing it to occur as a lockword, can be accounted for by its usefulness in building the narrative of the play into the dialogue of the characters.

The concordance data for both corpora shows that WORLD tends to be used in a hyperbolic way, not to describe the earth or the globe but to mean the "whole of mankind", as Crystal and Crystal (2002:503) state of its use in Shakespeare's plays.

Characters use expressions like *all the world*, *in the world* and *the wide world* to boost or exaggerate what they are saying, as shown in the following extracts, one from each corpus.

(19) Petruchio: Why does the world report that Kate doth limp?
O slanderous world!

Shakespeare, *The Taming of the Shrew*, II:i (*SDC*)

(20) Mistress Mary: If all the world else would forgive the deed,
Yet would I earnestly pursue the law.

Heywood, *How a Man May Chuse* (*NDC*)

In example (19) above, Petruchio exaggerates his teasing of Kate by claiming that the world is talking about her, and in example (20), Mistress Mary emphasises how strongly she feels about justice by distancing herself from what "all the world" would do. The lockword EVERY is used in a similar way, to boost or add emphasis to what speakers in both corpora say by exaggeration (usually through metaphor), as in the extracts from each corpus below.

(21) Titania: And when she weeps, weeps every little flower,

Shakespeare, *A Midsummer Night's Dream*, III:i (*SDC*)

(22) Ptolemy: Let stormy hail and thunder beat on him
And every bird and beast run over him,

Chapman, *The Blind Beggar of Alexandria* (*NDC*)

Characters strengthen the emphasis of what they are saying by adding "every": in example (21) it is not just flowers that weep when the woman Titania describes is weeping, it is *every* flower. In example (22), Ptolemy wants *every* creature to run over another character, not just one or two. The locking of WORLD and EVERY across both corpora indicates that they were popular in rhetorical strategies of exaggeration or

hyperbole which Shakespeare and other contemporaneous dramatists frequently made use of.

Finally among the lockwords in Table 17, one clearly has a function in stagecraft. FAREWELL, noted as a "formal" parting formula by Cusack (2004 [1970]:119), is a language strategy which enables characters to depart (so they and the plot can move on). The high-frequency content words COME and LET in both corpora, discussed in 6.2, were similarly instrumental in moving characters on and off stage. As Herman (1995:162) points out, "[e]xits require characters to leave the stage. Before exits can be performed, characters have to disengage from interaction." A link can be made here with Arnovick's historical sociolinguistic research. She states that in Shakespeare's plays *farewell* is much more frequent than other "parting salutations" such as *adieu* and *God be with you* (including its variant forms) (1999:96). This, Arnovick argues, gives "[a] sense of the relative frequency of the different closings in Early Modern English". The occurrence of *farewell* as a lockword in my data adds strength to Arnovick's claim, through showing a similar level of frequency of use in a collection of synchronic works by a range of other dramatists (and also relatively lower frequencies for *adieu* and forms of *God be with you*[54]).

Overall, the words which lock across the corpora have highlighted some specific ways in which Shakespeare and the other contemporaneous dramatists share preferences for similar language strategies in constructing character dialogue, over and above what the high-frequency words showed in 6.2. PARDON has been added to the deferential terms which are frequently used between characters, and FELLOW has

---

[54] Arnovick uses Spevack's (1969) Shakespeare corpus, in which the frequency counts are *farewell* (520); *adieu* (104) and *God be with you* (16) (Arnovick 1999:96). These are lower than the frequencies in the *SDC*, probably because the *SDC* is limited to just the plays in the First Folio. The frequencies in the other contemporaneous plays (in the *NDC*) are *farewell* (373); *adieu* (48) and forms of *God be with you* (8).

been uncovered as a non-deferential term which occurs frequently in both corpora. Similar ways of describing things have been identified, through MIND and SET, and similar ways of exaggerating, through WORLD and EVERY. Some of the overarching themes in EModE plays are also clearly visible through the words which lock across both corpora (MAJESTY, POWER and DEATH), adding to the theme of LOVE identified in the most frequent words in 6.2. In the next section, I look at the 3-word clusters which "lock" across the corpora.

7.3 Locked 3-word clusters

Table 19 shows the top 20 locked 3-word clusters, their raw frequencies in both corpora, and the p values.

Table 19. Top 20 rank-ordered locked 3-word clusters in Shakespeare's plays and other contemporaneous plays (minimum frequency=50; p=1.0)

Rank	Word	Shakespeare's plays	Other contemporaneous plays	p value
1	I AM SURE	69	72	0.987
2	FOR I AM	52	55	0.956
3	THOU ART A	78	83	0.915
4	I AM GLAD	58	62	0.907
5	GIVE ME	50	51	0.898
6	OF THE WORLD	52	56	0.881
7	I HAVE NO	57	58	0.880
8	MY LORD OF	97	99	0.862
9	AND ALL THE	79	80	0.836
10	I HAVE NOT	57	63	0.764
11	WITH ALL MY	51	57	0.732
12	I WILL NOT	213	215	0.710
13	I AM A	139	138	0.662
14	AND IN THE	57	65	0.635
15	BUT I WILL	55	63	0.624
16	I DO NOT	157	155	0.608
17	NOT TO BE	54	50	0.533
18	I HAVE A	97	90	0.411
19	I AM THE	50	44	0.402
20	I HAVE BEEN	66	57	0.288

Table 19 shows that the most strongly locked cluster is one through which characters express certainty: I AM SURE. As with the top-ranking lockword FELLOW, in the previous section, this is in some ways surprising, since it does not appear (to me) to be psychologically prominent in the play-texts. However, unlike FELLOW, it has been a focus of analysis in another corpus study of literary texts, as I explain below. The dispersion plots from my data show I AM SURE to be distributed fairly evenly across all three genres in the other contemporaneous plays, although rare in Shakespeare's history plays compared to the comedies and tragedies. A selection of examples from the concordance data is given below in Figures 9 and 10, from the *SDC* and the *NDC* respectively.

N	Concordance	File
1	en that Dobbin's tail grows backward: I am sure he had more hair on his ta	1\scmov.txt
2	Neighbour, this is a gift very grateful, I am sure of it. To express the like kin	\scshrew.txt
3	Then you say as I say; for I am sure he is not Hector.	\sctandc.txt
4	las!' I would fain say, bleed tears, for I am sure my heart wept blood. Who	\scwtale.txt
5	, Kate? I will tell thee in French, which I am sure will hang upon my tongue li	1\shhenv.txt
6	Why, so he did, I am sure. No, no;	1\stcorio.txt
7	our lady does not love her husband; I am sure of that: and at her late bein	~1\stlear.txt
8	w not that; but such a handkerchief- I am sure it was your wife's-did I toda	r~1\stoth.txt
9	e him. So will ye, I am sure, that you love me.	1\strandj.txt
10	st men. Ye've heard that I have gold; I am sure you have: speak truth; ye're	1\sttimon.txt

Figure 9. Concordance extract for *I am sure*: Shakespeare's plays

N	Concordance	File
1	would drive away the time trimly, come I am sure you are not without a score.	~2\ncoldwi.txt
2	hs near to my house, They are not far I am sure, if I make haste,	\nctwoang.txt
3	All these Are out of hope, I am sure the man.	r~2\ncvolp.txt
4	ay, that bears to landward, That way, I am sure they will not take, Go make	~2\nhduch.txt
5	Call you this a little thundering, I am sure my breeches finds it a great	ar~2\nhedi.txt
6	King's Butler, and Tom of his Chamber, I am sure ye know them?	~2\nhedivi.txt
7	pray you sir, let her go along with us, I am sure his honour will welcome her,	ar~2\ntaof.txt
8	pay. you wonder I am sure whence this strange kindnes	r~2\ntawkk.txt
9	. From them, I should learn somewhat I am sure I never shall know here: I'll t	r~2\ntdom.txt
10	s I came home, he slipped me in, And I am sure he is with Abigall.	ar~2\ntjew.txt

Figure 10. Concordance extract for *I am sure*: other contemporaneous plays

Interestingly, Fischer-Starcke (2010:120-127) finds that *I am sure* is a frequently-occurring phrase in her corpus of Jane Austen's 19th century novels. Fischer-Starcke

(2010:123) states that it "functions as an eye catcher to attract the reader's attention and to direct it to that part of the utterance which is most relevant for its meaning." She argues that this is an authorial style feature rather than a register feature of literary texts, because the phrase does not occur in Stubbs and Barth's (2003) fiction corpus (2010:123). However, its prevalence in my EModE drama data by Shakespeare and other contemporaneous playwrights indicates that Austen is not the first to make use of it as a strategy for alerting the reader/audience to something important. From analysing its use by speakers in one novel, *Northanger Abbey*, Fischer-Starcke (2010:124) also finds that *I am sure* "contributes to characterizing both the place Bath and its society as superficial". However, this does not seem to be an effect created by the phrase in my data from EModE dramatic dialogue of several centuries earlier, because the concordance data does not show any evidence that *I am sure* is associated with people in particular places. The kinds of characters who habitually use *I am sure* would be worth exploring further in future research, though, to see if it is associated with speakers and/or addressees of particular social rank(s) or gender in EModE plays.

Further down Table 19, more than half of the clusters which lock across both corpora are in the first person and feature the pronoun *I*, which was identified as a lockword in the previous section. The locked clusters provide more information about the formulaic ways in which characters in both corpora tend to talk about themselves. To see more clearly the functions of the 3-word locked clusters, especially the self-oriented ones featuring *I*, I now categorise them according to Culpeper and Kytö's (2010:107-111) framework (in the same manner as the high-frequency clusters in 6.3). The functional categories of the top 20 locked clusters are shown in Table 20 on the next page, with the raw frequencies given for each corpus in brackets alongside (the first figure is for the *SDC* and the second is for the *NDC*).

Table 20. Functions of top 20 3-word clusters which lock across Shakespeare's plays and other contemporaneous plays

Interpersonal	Speech act-related	*Directive* GIVE ME LEAVE (50, 51)
	Modalising	*Volition* I WILL NOT (213, 215) *Shield/certainty marker* I AM SURE (69, 72) WITH ALL MY (51, 57) *Approximator/intensifier* OF THE WORLD (52, 56) AND ALL THE (79, 80)
Textual	Organisational	*Informational elaboration* FOR I AM (52, 55) AND IN THE (57, 65) BUT I WILL (55, 63)
Ideational	Topical	*People* MY LORD OF[55] (97, 99) *States* THOU ART A[56] (78, 83) I AM GLAD (58, 62) I HAVE NO (57, 58) I HAVE NOT (57, 63) I AM A (139, 138) I DO NOT (157, 155) NOT TO BE (54, 50) I HAVE A (97, 90) I AM THE (50, 44) I HAVE BEEN (66, 57)

Table 20 shows that half of the top 20 clusters which lock across both corpora are those which describe states (in the Ideational: Topical category), particularly states of being, doing or having. All but two of the States clusters (THOU ART A and NOT TO

[55] MY LORD OF is a term of reference in most cases in the data, e.g. *my lord of York*, not a term of address; in other contexts, *my lord* has a vocative function in the dialogue in both corpora.

[56] Though THOU ART A contains an address form, the principle of classifying clusters according to their highest discourse function (explained in 3.2.4, following Culpeper and Kytö 2010:111) means it is classified as describing a state. Notably, this cluster occurs in speech acts of complimenting or insulting an addressee in the play-texts (e.g. *thou art a sweet lady, thou art a fool*).

BE) are self-oriented and, as with the high-frequency cluster data when the corpora were examined independently of one another (in 6.3), they serve a primary dramatic purpose of describing the people and the activities in the plays. A quarter of the locked clusters fall into Modalising categories, including those which express volition, certainty or which otherwise intensify what characters are saying. The certainty markers I AM SURE and WITH ALL MY convey strength of feeling and emotion; most instances of the latter are embedded in the longer formula *with all my heart*. The intensifying clusters OF THE WORLD and AND ALL THE exaggerate what is being said. AND ALL THE is in almost all cases a generalisation (e.g. in the expression *and all the world*). In the previous section I discussed the hyperbolic use of the lockword WORLD, and now the locked cluster results add to the evidence that references to the whole world are rhetorical strategies used relatively frequently by Shakespeare and other dramatists of the same period. Taken together, the Modalising locked clusters convey strength of feeling and emotion in various ways: i.e., they help create the sense of drama itself.

As with some of the high-frequency clusters discussed in 6.3 (with reference to Culpeper 2001:167-172), some of the locked clusters in Table 20 have functions of self- and other-presentation in characterisation. FOR I AM, I AM GLAD, I AM A, I AM THE are used by characters to present some aspect of themselves, and THOU ART A is used to present some aspect of another individual's character. The other first-person clusters in Table 20 also arguably contribute to self-presentation too, although the link seems strongest in those featuring forms of the verb *to be*.

It is worth noting that although the word *not* does not arise as a lockword on its own (in 7.2 above), it does occur in a number of locked clusters which are negated (the Volition cluster I WILL NOT and the States clusters I HAVE NOT, I DO NOT and

NOT TO BE). In discussing the high-frequency *not* clusters in 6.3, I argued that negation contributes to characterisation and helps heighten the sense of dramatic excitement or tension, through characters describing themselves and others via strategies of denial or rejection. The concordance data shows similar effects created by the negated locked clusters, so I will not discuss them in further detail here.

The Vocative cluster MY LORD OF arguably results from the "aboutness" of the texts (discussed in 2.7), since it reflects the high proportion of noble characters in many of the plays in both corpora through its function of deference in address and reference to other characters.

Having examined words which lock at the lexical level in the previous section, and at those which lock at the lexico-grammatical level in this section, I now move on to the semantic level to see what kinds of concepts lock across both corpora.

7.4 Locked semantic domains

Table 21 on the next page shows the raw frequencies and log-likelihood values of the top 10 semantic domains which lock across the *SDC* and the *NDC*. These are the domains which occur with the most similar frequencies, i.e. those which are nearest to a p value of 1.0/log-likelihood ("LL") of 0[57]. The three most frequently-occurring words in each domain are shown as examples from each corpus, in italics, below the semantic category label.

[57] As noted in 2.5.2 and 3.4.1, *Wmatrix* computes only the log-likelihood value for results, not the p value.

Table 21. Top 10 rank-ordered semantic domains which lock across Shakespeare's plays and other contemporaneous plays (minimum frequency=200; LL=0)

Rank	Semantic domain	Shakespeare's plays	Other contemporary plays	LL
1	Dislike (E2-) e.g. *hate, beast, hateful (SDC)* e.g. *hate, beast, beasts (NDC)*	911	865	0.00
2	Weather (W4) e.g. *wind, clouds, rain (SDC)* e.g. *wind, clouds, thunder*	858	819	0.00
3	Clothes and personal belongings (B5) e.g. *crown, wear, suit (SDC)*	2,469	2,346	0.01
4	Furniture and household fittings (H5) e.g. *bed, gates, throne (SDC)*	828	792	0.01
5	Objects generally (O2) e.g. *spoke, stones, spring (SDC)*	2,541	2,415	0.01
6	Quantities: many, much (N5+) e.g. *many, much, enough*	2,083	1,990	0.01
7	Unlikely (A7-) e.g. *doubt, impossible, doubtful (SDC)*	306	288	0.02
8	General and abstract terms (A1) e.g. *thing, things, stuff (SDC)*	747	705	0.03
9	In power (S7.1+) e.g. *lord, sir, king (SDC)* e.g. *sir, lord, king (NDC)*	13,763	13,134	0.03
10	Farming and horticulture (F4) e.g. *shepherd, bred, breed (SDC)*	529	510	0.04

The semantic domains which lock statistically across both corpora are very different from the ones which arose most frequently when the *SDC* and the *NDC* were analysed independently. Those results (in 6.4) were populated mostly by categories of words relating to interactional speech and to the register of drama. In contrast, Table 21 shows that half of the domains which occur with the most similar frequency across the corpora encompass concepts of routine daily life: the weather, clothes and belongings,

objects, quantities, furniture, farming and horticulture. At first glance, the most striking effect is that most of them describe fairly mundane aspects of life. This is interesting, and the reasons behind it become clear below with some further exploration and discussion of the concordance data (along with a few problems in the categorisation of words by the USAS tool, as were noted among the high-frequency domains in 6.4).

The top two semantic domains in Table 21 rank equally for the greatest statistical strength (with the same log-likelihood value of 0). These are the most strongly locked domains, containing concepts of "Dislike" and "Weather". Looking first at the "Dislike" domain, by far the most frequent word in both corpora is *hate*. The word *beast*, also very frequent in both corpora, is slightly problematic because it is not always an expression of dislike, but apart from that the category is reliable, and contains words of passionate and extreme dislike, including *loathe*, *detest* and *despised*. Some examples are shown in the concordance extracts from the *SDC* (Figure 11, below) and the *NDC* (Figure 12, on the next page).

1.	n you do me greater harm than	hate	? Hate me ! wherefore ? O me !
2.	o me greater harm than hate ?	Hate	me ! wherefore ? O me ! what n
3.	r : 't is no jest, That I do	hate	thee and love Helena. O me !
4.	om these that my poor company	detest	: And sleep , that sometimes s
5.	eeking sweet favours for this	hateful	fool , I did upbraid her and f
6.	ve the boy, I will undo This	hateful	imperfection of her eyes : And
7.	o pass? O! how mine eyes do	loathe	his visage now . Silence , awh
8.	e concord in the world, That	hatred	is so far from jealousy , To s
9.	r from jealousy, To sleep by	hate	, and fear no enmity ? My lord
10.	ut , like in sickness, did I	loathe	this food ; But , as in health

Figure 11. Concordance extract for semantic domain E2- (Dislike): Shakespeare's plays

1.	beast my man Manes. He is a	beast	indeed that will serve thee .
2.	or birth , and then as good	hated	as enforced . I am a king , a
3.	needs twice perish with his	hate	, and thine own love . Thy pa
4.	easure of peace , unless you	despise	the rudeness of war . It is s
5.	little worser than I can of	hate	. And why ? Because it is bet
6.	ngs , which give occasion of	hate	. Why , be not women the best
7.	en and Bees . What dost thou	dislike	chiefly in a woman ? One thin
8.	unatike as men suppose , but	hates	company , and worldly trash ,
9.	yea and perhaps lousy , with	despising	the vain shifts of the world
10.	dearly as I do , to make you	hateful	in his sight , that I might m

Figure 12. Concordance extract for semantic domain E2- (Dislike): other contemporaneous plays

In my data, hating is most often used to express the very antithesis of liking or love, associated with people or events. Tissari's (2010b) findings, which are mainly from the register of EModE correspondence, indicate that there may be implied religious beliefs and values underpinning who is hated, and in what prompts or fuels hating. Furthermore, hating might not have been evaluated in a negative way: Tissari (2010b:311) comments that "while 'hate' is a bad thing as such, it is nevertheless acceptable, even desirable, to hate that which is bad, in a religious sense [...] In doing so, we identify with God, who hates sin." Although my data is from a different register, it nevertheless seems likely that the considerable authority of religion in society at this time would have influenced the evaluation and interpretation of hating in the plays by contemporaneous audiences.

Regardless of the connotations of hating which may have been perceived by a contemporaneous audience, however, the word *hate* and others in the "Dislike" domain in my data are useful vehicles for conveying some strong emotional forces which are very prevalent in plays by Shakespeare and his peers, such as revenge, ambition for power, and/or love. There was some evidence of these as common themes in the corpora in the lockwords DEATH and POWER, in 7.2 above. (Further down Table 21, "In power" appears as a domain which locks across both corpora, and is mainly populated with terms reflecting hierarchical social structures, such as

deferential vocatives, e.g. *sir* and *lord*, and words such as *command*, *lead*, *overthrow*, *sovereignty*, *enforce* and *power*.) However, the strongly locked "Dislike" domain adds to what the lockwords show, since it brings to the fore some evaluative language which helps to contextualise the recurrent themes of power, death, and associated themes such as revenge and jealousy. The extreme emotional reactions of dislike conveyed in the characters' language overlay the major themes in plays with a sense of drama. The voicing of characters' emotional reactions helps dramatise what would otherwise be just a set of events narrated by the characters. The expression of extreme dislike accents the events of the plays with passion and feeling.

Since strongly emotional language associated with extreme dislike would seem to have a role in conflict, either in the build-up to it, or as a result of it, I expected that language in the "Dislike" domain would tend to be concentrated more heavily in the early to middle stages of plays in the corpora. This is based on arguments regarding typical plot shapes and points of conflict. Culpeper (1997:87) cites Bremond's (1966, 1973) view that dramatic plots tend to progress from an initial state of equilibrium, to disequilibrium, and back to equilibrium again, with states of disequilibrium being characterised by conflict. Hapgood (2004 [1967]:147), who analyses "modes of speech" in some of Shakespeare's history plays, argues that *Richard II*, both parts of *Henry IV*, and *Henry V* follow a shape of "initial disorder through virtual chaos to a final restoration of order". As explained in 3.5, the distribution of semantic domain output from *Wmatrix* can be plotted in *AntConc* by downloading the data file and searching on the tag label. The distribution of the "Dislike" semantic domain in the four Shakespearean plays mentioned by Hapgood is shown in Figure 13 on the next page.

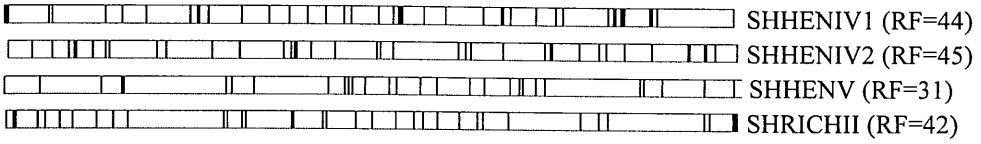


Figure 13. Distribution of "Dislike" concepts in 4 Shakespearean history plays

Figure 13 shows that language associated with disliking is actually concentrated near the end of *Richard II* and both parts of *Henry IV*, as well as in earlier stages. Only *Henry V* shows a distribution pattern which could be said to fit with the recession of conflict in later stages of the play. Since the general distribution of items in the "Dislike" domain did not show any compelling evidence for points of conflict that accord with arguably typical plot shapes in the plays in my corpora, I narrowed the investigation to just the most frequent word in the domain in both corpora: *hate*. This is a very specific concept of extreme dislike, through which characters reveal their feelings and motivations (examples of which are shown in Figs. 11 and 12 above).

The distribution of *hate* is shown for the *SDC* in Figure 14, and for the *NDC* in Figure 15, on the next page. The plays in each set of data are listed in order of comedies, histories and tragedies, the genre being indicated by the second letter of the filename (as explained in 5.2; see further Appendix IV for a full list of text-ids for the plays in each corpus).

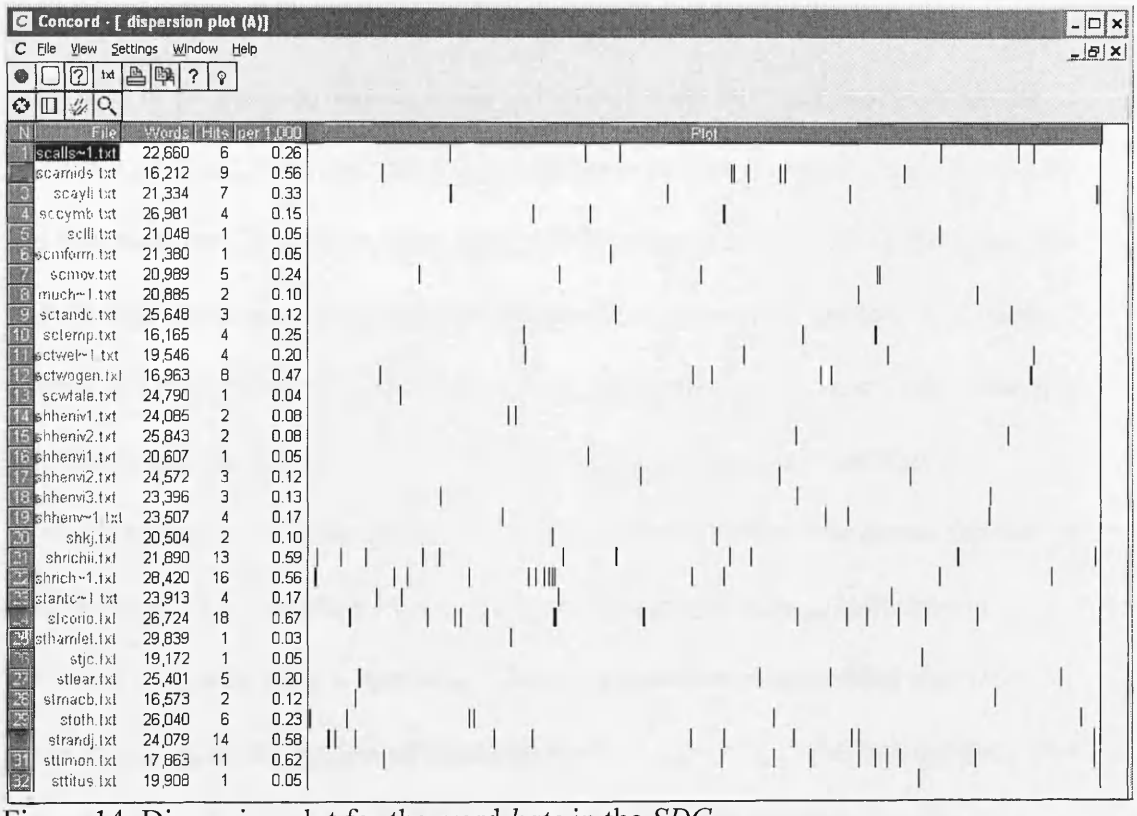


Figure 14. Dispersion plot for the word *hate* in the *SDC*

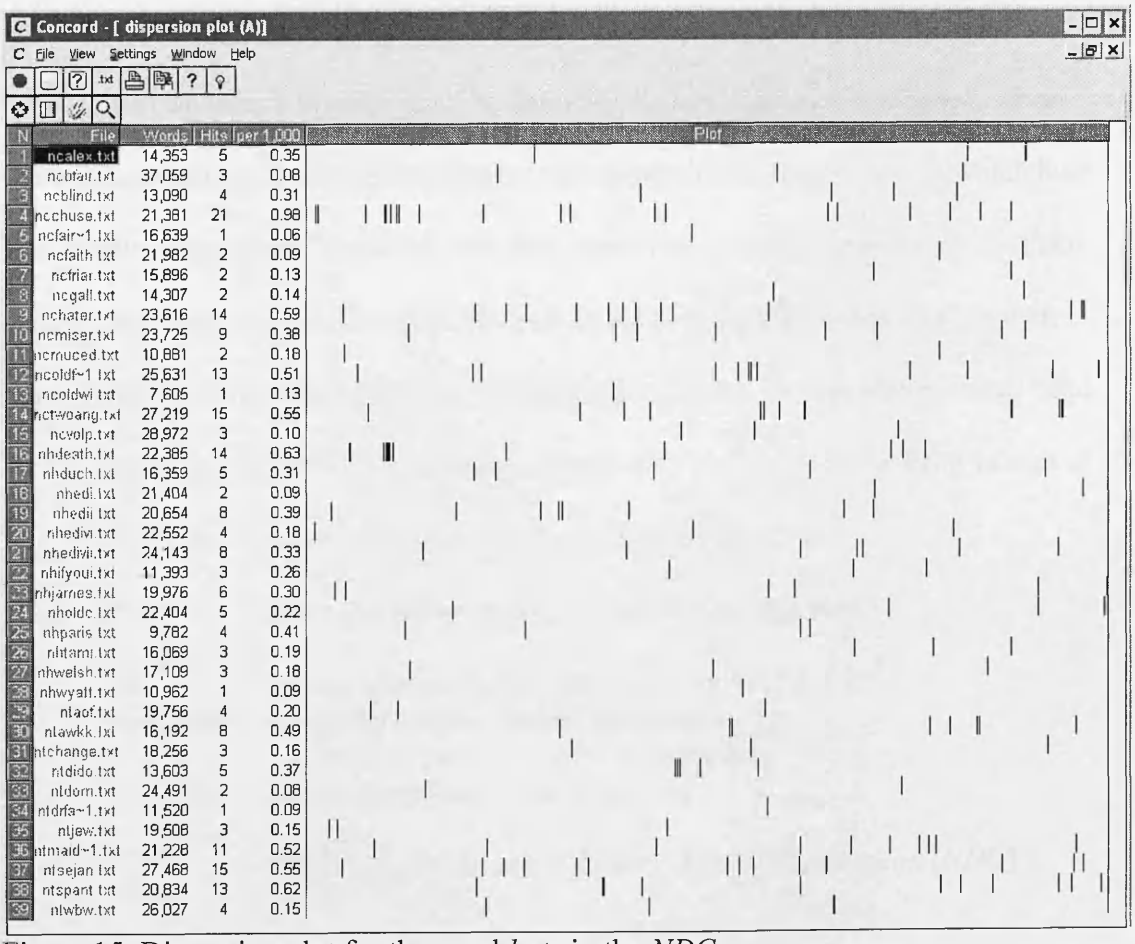


Figure 15. Dispersion plot for the word *hate* in the *NDC*

In Figure 14, a minority of Shakespeare's plays show a visible concentration of the word *hate* in the early or middle stages compared to the later stages. These are the comedies *As You Like It* and *Cymbeline*, the history plays *Richard II* and *Richard III*, and the tragedies *Coriolanus*, *King Lear* and *Romeo and Juliet*. Not all the tragedies feature high incidences of *hate* (*Macbeth* and *Titus Andronicus* are low, despite the violent acts they contain). This is interesting, suggesting that there are other concepts apart from extreme forms of dislike which construct situations of conflict in Shakespeare's plays. Figure 15 shows a heavier distribution of *hate* across the early to middle stages of fewer plays by Shakespeare's contemporaries. It is evident in Heywood's comedy *How a Man May Choose*, where *hate* is something characters talk about avoiding, in the context of discussions about choosing a marriage partner. This is topical, though, rather than instrumental in precipitating conflict.

The strongest concentration of *hate* in the early stages of a play, in either corpus, is in Munday's history play *The Death of Robert, Earl of Huntingdon*, where there is clear evidence that it functions in conveying character motivations which lead to a middle stage of conflict in the plot. The word *hate* is used repeatedly by the Prior and Sir Doncaster in a conversation early on in the play, in which they explain their reasons for wanting Robert (the Earl of Huntingdon, also known as Robin Hood) dead. Sir Doncaster envies Robert's popularity compared to his own, and the Prior is next in line to inherit the earldom. An extract is shown in example (23).

(23) Prior: But tell me Doncaster, why dost thou hate him?

Sir Doncaster: By the Mass, I cannot tell. O yes, now I have it.
I hate thy cousin, Earl of Huntington,
Because so many love him as there do,
And I myself am loved of so few.

Munday, *The Death of Robert, Earl of Huntingdon* (*NDC*)

Example 23 supports my expectation that language associated with "Dislike" would appear to be linked with setting up states of conflict, and thereby a driver of plot shape. However, it is an exception, not typical of all the plays in the corpora. This suggests that language from other semantic fields performs the same function in other plays (from the scarcity or absence of results in plays with known elements of conflict and malignant dislike, such as Shakespeare's *Titus Andronicus*). Therefore, there seems not to be a clear link between language expressing dislike and the plot shapes indicated by Bremond (1963, 1973) and Hapgood (2004 [1967]:147). Shakespeare and his contemporaries make use of language associated with extreme dislike even in the resolution stages of plays, and its function of adding drama, discussed above, is more likely to explain its prevalence in both sets of EModE plays. However, the examination of the distribution of a semantic domain, in whole or in specific part, is a potentially informative and useful analytical step.

I turn now to the other most strongly locked semantic domain in Table 21, "Weather". The concordance data shows that weather concepts are much more associated with inclement conditions than clement ones, in both corpora (e.g. rain, clouds, thunderbolts, lightning, storms, torrents, floods and storms). Some examples are shown in the extracts from the *SDC* (Figure 16, below) and the *NDC* (Figure 17, on the next page).

1.	season , For thou mayst see a	sunshine	and a hail In me at once ; but
2.	ou mayst see a sunshine and a	hail	In me at once ; but to the bri
3.	he brightest beams Distracted	clouds	give way : so stand thou forth
4.	ellent young man ! If I had a	thunderbolt	in mine eye , I can tell who s
5.	lish chiding of the winter 's	wind	, Which , when it bites and bl
6.	No enemy But winter and rough	weather	. More , more , I prithee , mo
7.	No enemy But winter and rough	weather	. I 'll give you a verse to th
8.	l , as large a charter as the	wind	, To blow on whom I please ; f
9.	g . Blow , blow , thou winter	wind	, Thou art not so unkind As ma
10.	riends ; that the property of	rain	is to wet , and fire to burn ;

Figure 16. Concordance extract for semantic domain W4 (Weather): Shakespeare's plays

1.	pray for them . The violent	thunder	is adored by those Are pasht
2.	ble wishes , I am prompt As	lightning	to your service , O my Lord
3.	eyes , or call her brow the	snow	of Ida , or Ivory of Corinth
4.	ue there arose me thought A	whirlwind	, which let fall a massy arm
5.	ed a fearful and prodigious	storm	, Be thou the cause of all e
6.	nt her . Let 's not talk on	thunder	, Thou hast a wife , our sis
7.	wonder much , What amorous	whirlwind	hurried you to Rome Devotion
8.	rother rage Beyond a horrid	tempest	or sea-fight , My vow is fix
9.	letoe on sear Elms spent by	weather	, Let him cleave to her and
10.	umy , Shipwrecks in Calmest	weather	? What are whores ? Cold Rus

Figure 17. Concordance extract for semantic domain W4 (Weather): other contemporaneous plays

While some are literal references, weather concepts are often used in metaphorical contexts (e.g. lines 1 and 2, Fig. 16; line 4, Fig. 17). Distribution plots for weather concepts in the *SDC* and the *NDC* are shown in Figures 18 and 19, respectively, below. They are annotated with labels indicating the plays in each corpus in which the greatest concentrations of weather concepts are found.

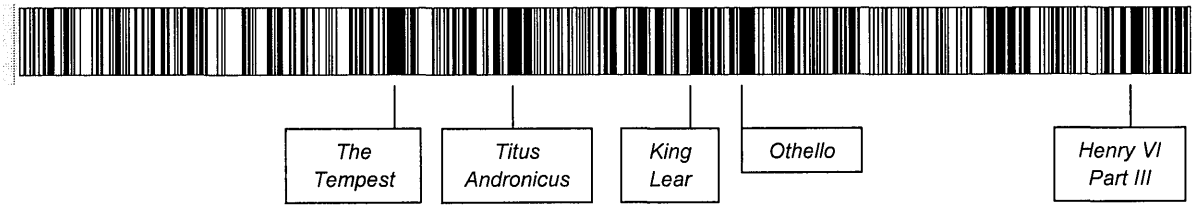


Figure 18. Distribution plot for weather concepts (USAS tag W4) in the *SDC*

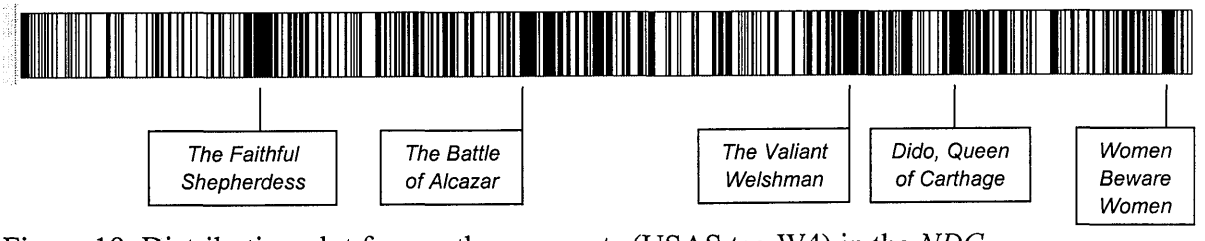


Figure 19. Distribution plot for weather concepts (USAS tag W4) in the *NDC*

Figures 18 and 19 show that although weather concepts are broadly distributed across the whole of each corpus, there are a few particularly dense concentrations in certain plays. In Shakespeare's plays (Fig. 18), it is perhaps not surprising that one of these is the comedy *The Tempest*, whose title implies that weather is a central theme. Interestingly though, the word *tempest* itself only occurs four times in the play, which

is no more than in some of the others (it occurs five times in Shakespeare's *Henry VI Part II* and five times in Webster's *The Duchess of Malfi*, for example). Three of the strongest concentrations in the Shakespearean data are in tragedies (*Titus Andronicus*, *King Lear* and *Othello*), and the other is in a history play (*Henry VI Part III*). In the other contemporaneous plays (Fig. 19), again there is one concentration in a comedy (Fletcher's pastoral comedy *The Faithful Shepherdess*), two in history plays (Peele's *The Battle of Alcazar* and Armin's *The Valiant Welshman*), and two in tragedies (Marlowe's *Dido, Queen of Carthage* and Middleton's *Women Beware Women*).

The connotations of weather concepts in the two sets of plays are various and interesting: my data indicates that sunshine has positive connotations, but conditions such as wind and cloud can be either positive, negative or neutral. Wind can describe a force for change, and cloud a sense of opacity or lack of clarity. Thunder can describe strength or loudness, and metaphorical rain sometimes has negative connotations but is also a way of describing the force or distribution of something (e.g. *rain down*). The extract in example (24) contains the first three weather references from the concordance data in Figure 16 above, through which the King describes his forgiveness of Bertram in Shakespeare's comedy *All's Well That Ends Well.*

(24) King: I am not a day of season,
For thou mayst see a sunshine and a hail
In me at once; but to the brightest beams
Distracted clouds give way: so stand thou forth;
The time is fair again.

Shakespeare, *All's Well That Ends Well*, V:iii (*SDC*)

In example (24), the King describes his change of disposition towards Bertram from bad to good as "hail" and "sunshine", respectively, and expresses his forgiveness as the clearance of "distracted clouds". The King is influenced by his good feelings (the sun, or "brightest beams") and invites Bertram to come before him, signalling their

reconciliation through another phrase that can also be applicable to weather ("fair again"). In his dictionary of the context of Shakespeare's plays, Richmond (2002) argues that weather was considered to have an important influence on people's behaviour as well as on their environment in the Early Modern period. Richmond states that:

> weather conditions, the state of society and individual temperaments are carefully presented in the plays as interactive, mutually dependent and congruous. Each play's stage is set astrologically, seasonally and weather wise to match its theme, characters and mood. [...]
>
> In Shakespeare's plays, weather conditions are usually established by descriptive speeches. This meteorology affects his characterizations. (2002:online edition[58])

This helps to explain the high frequency with which weather concepts occur in my Shakespearean data. The fact that weather concepts are so strongly locked across both corpora indicates that other contemporaneous dramatists found it to be of similarly high importance in constructing characters, events and actions through dialogue. The weather is often personified in both sets of plays, as in the following example from Jonson's tragedy *Sejanus*.

(25) Sejanus: You know, sir, thunder speaks not till it hit.

Jonson, *Sejanus*, II (*NDC*)

In example (25), thunder (a non-animate entity) is accorded the power of a speaking voice (a human quality and ability), creating an effect of personification described as "pathetic fallacy" by scholars in literary and linguistic disciplines. The stylisticians Leech and Short (2007) include pathetic fallacy in their discussions of mind style in prose fiction, noting a range of linguistic guises in which it appears. In their analysis of an extract from Thomas Hardy's novel *The Return of the Native*, they argue that:

---

[58] Accessed online at http://www.credoreference.com/entry/contst/weather (18.01.12).

> In its weakest and least specific form, the pathetic fallacy is manifested in the use of inanimate nouns as actors, for example as implied subjects of verbs of motion [...] More specific are expressions which attribute motive and feeling to inanimate nature (2007:159)

Richmond (2002) illustrates the use of pathetic fallacy in Shakespeare's plays with a weather metaphor from the dialogue of *King Lear* (III:ii), and he argues that:

> the pathetic fallacy is an intrinsically dramatic mode of bringing non-human factors into expression in a script without the possibility of full representation of **storms**, which cinematic techniques now permit (2002:online edition, Richmond's emphasis)[59].

Richmond's explanation suggests that pathetic fallacy is used in EModE drama by way of creating special effects in the mind of the audience, through language rather than through the audio-visual means which are familiar to present-day audiences of television and film drama. In my data from the "Weather" semantic domain alone, there are numerous instances of the kinds of pathetic fallacy described by Leech and Short (2007) and Richmond (2002), some of which are evident in the examples above. In the King's speech from *All's Well That Ends Well* (example (24)), the description of his moods as "sunshine" and "hail" is a weak form of pathetic fallacy, because the inanimate nouns are incorporated into a sense of human feeling or emotion but they do not perform any actions. The "distracted clouds" in example (24), and the thunder in example (25) from *Sejanus*, are both stronger forms, as they are accorded the power of movement and speaking, respectively, and the motivation to do so. From their analysis of the Hardy extract, Leech and Short conclude the following about pathetic fallacy:

> The personifying metaphor is so consistently employed that 'metaphor' almost ceases to be the appropriate term: it is as if our literal sense of the division between animate man and inanimate nature has been eliminated. (2007:160)

---

[59] Accessed online at http://www.credoreference.com/entry/contst/pathetic_fallacy (18.01.12). Richmond (2002) notes that the term "pathetic fallacy" originated with John Ruskin, who used it as a pejorative description.

Their observation suggests that the overall effect of pathetic fallacy, when it is skilfully deployed by an author, is to bypass what might be a conscious or partly-conscious distinction between the literal and the metaphorical on the part of readers (and, by extension, audiences). This would help achieve a state of receptiveness to the vivid, creative and exaggerated descriptions which make prose fiction and drama exciting and entertaining, but which would seem ridiculous or at the very least unconventional in other language contexts. In drama, special effects created by language seem plausible when presented through pathetic fallacy, in a similar way that the magic of the screen makes audio-visual special effects convincing today. The effects seem plausible because the audience cannot see exactly how they are achieved, and so they are able to take in exaggerated or alternative forms of reality with which they are presented, and become engaged with the drama.

Although references to weather in plays by Shakespeare and by other contemporaneous dramatists are notably highly frequent, they do not seem to have been the focus of other studies which concern metaphor in EModE drama (discussed in 3.3.3). In Tissari's (2009) analysis of the concept of the soul, only one item of her data from Shakespeare's plays includes a weather concept (a storm, in her example number 37). Weather concepts are worth investigating further in EModE drama by Shakespeare and other dramatists, in a future study, to explore the kinds of metaphorical contexts they are most often used in. The contents of the "Weather" semantic domain were generally reliable, apart from the word *rime* being inappropriately included, as a result of EModE spelling conventions. Although the spelling in the corpora was regularised using *VARD 2*, as explained in 5.4, this did not extend to the modernisation of all the EModE spellings. Although the *OED* confirms that this word form has a meaning associated with frost as far back as the Early

Modern period, all the instances in my corpora have the semantic meaning of the PDE word form *rhyme* (e.g. *rime and reason*). As there are only 42 instances of *rime* out of a total of 858 lexical items relating to weather (across both corpora), it is unlikely to affect the overall strength of weather as a locked domain.

The semantic domains in Table 21 which encompass aspects of everyday social life are generally interesting for the snapshot they provide of personal, household and domestic concepts which are mentioned frequently by Shakespeare and the other dramatists. Many of these are literal references (e.g. to clothing, furniture and other everyday objects), while others are used in metaphors. Some of the latter fit with the arguments of Lakoff and Johnson (1980:3), mentioned in 3.3.3, that conceptual metaphors are characteristic of the language of routine social life. Though EModE drama features some central themes (such as power and revenge, discussed above and in the previous sections), these are fleshed out by dialogue which describes the objects, events and activities of people's lives and surroundings in the plays, such as what they wear, use and do in their homes. I discuss some examples briefly below, together with a few problems with categorisation by the USAS tool. The concordance extracts in Figures 20 and 21 on the next page show examples from the "Clothes and personal belongings" domain, in the *SDC* and the *NDC*, respectively.

1.	! A sentence is but a cheveril	glove	to a good wit : how quickly the
2.	, would you undertake another	suit	, I had rather hear you to soli
3.	s many lies as will lie in thy	sheet	of paper , although the sheet w
4.	and will laugh yourselves into	stitches	, follow me . Yond gull Malvoli
5.	of grossness . He 's in yellow	stockings	. And cross-gartered ?
5.	t me . Hold , sir ; here 's my	purse	. In the south suburbs , at the
6.	shall you have me . Why I your	purse	? Happily your eye shall light
7.	n my master 's griefs . Here ;	wear	this jewel for me , 't is my pi
8.	at 's certain , or forswear to	wear	iron about you . This is as unc
9.	orribly conceited of him ; and	pants	and looks pale , as if a bear w
10.	ale , as if a bear were at his	heels	. There 's no remedy , sir : he

Figure 20. Concordance extract for semantic domain B5 (Clothes and personal belongings): Shakespeare's plays

1.	ease a woman that like a dutch	doublet	all his back is shrunk into his
2.	ll his back is shrunk into his	breeches	. Shroud you within this closet
3.	he comes , this fellow by his	apparel	Some men would judge a politici
4.	t like a hound In Leon at your	heels	. Tweare for her honour And so
5.	me you know not where my	night cap	wrings me . Wear it ath' old fa
6.	where my night cap wrings me .	Wear	it ath' old fashion , let your
7.	ut one twelve pence a th' bord	twill	appear as if there were twenty
8.	courtly gentleman , - when he	wears	white satin one would take him
9.	tleman , - when he wears white	satin	one would take him by his black
10.	leaves none What value is this	Jewel	It is the ornament Of a weak fo

Figure 21. Concordance extract for semantic domain B5 (Clothes and personal belongings): other contemporaneous plays

The importance and significance of attire in the Early Modern period, and what it represents on-stage, is noted in other research, for example by Burkert (2011), Jardine (1983:141-168) and Orgel (1996). Jardine states that "[d]ress, in the early modern period, was regulated by rank, not by income" (1983:141), so that even those with sufficient means could not wear whatever they wanted unless they were of sufficiently high social status. Jardine provides a table showing what kinds of cloth, colours and embellishments were permitted for persons at each level of social status (1983:143-144). However, in a time of "uneasy relationships between old nobility and hereditary titles on the one hand, and the newly enriched on the other" (Hattaway 2003:99), the enforcement of dress codes by law became contentious, and Orgel (1996:98) states that the laws governing dress and social status were actually dispensed with in 1604. This, Orgel argues, "transferred the jurisdiction over questions of appropriate dress, as

an issue of public morality, from the criminal courts to the ecclesiastical ones, where the guidelines were much less clear" (1996:98). The crux of the matter of dressing appropriately to social status was anxiety over "social imitation" (Orgel 1996:100), i.e. the problem of dressing in a manner above one's station. The fixed association between dress and social rank that existed in the Early Modern period meant that clothes themselves became "legitimating emblems of authority", according to Orgel (1996:105), and there was a sense in which a person's social identity could be transformed by what he or she wore. Orgel (1996) discusses the implications for this with regard to the audience's reception of boy players dressed in women's costumes for female roles in the plays, something that there is regrettably no space to discuss here.

The discussions of Jardine (1983) and Orgel (1996) make clear the potential for dramatists to criticise, comment upon and/or satirise issues of dress and social behaviour in plays, and also the possible social implications in dramatic dialogue which makes reference to clothing. The full extent of these implications and subtle ironies may well go unnoticed by a present-day reader/audience, though the basic meanings are still clear. I will illustrate this with the complete line from which the first concordance extract in Shakespeare's plays is taken (in Fig. 20 above), which is shown in example (26).

(26) Clown: A sentence is but a cheveril glove to a good wit:
how quickly the wrong side may be turned
outward!

Shakespeare, *Twelfth Night*, III:i (*SDC*)

The analogy of the glove (like a sentence, being able to be turned inside out by someone who has the dexterity) is still clear today, even without the knowledge that *cheveril* means "kid-leather", and is a word which, according to Onions (1983:34), is "always used allusively as a type of flexibility" in Shakespeare's plays. There are only

three instances of *cheveril* in the *SDC*[60], so on the one hand it would be unlikely to arise as a result on its own on the basis of frequency, and is therefore statistically relatively unimportant. On the other hand, it does not occur in the *NDC* at all, and is therefore of stylistic interest as a creative way of describing flexibility which Shakespeare made use of, but which a range of his peers apparently did not. (The *NDC* contains but a selection of all EModE plays, though, and I have not examined all the other works of the dramatists whose work it includes.) The "Clothing" domain provides much scope for interesting future research, particularly with regard to gender and social rank of characters. The important point here is that a potentially interesting style feature has surfaced through being grouped together with other words into a larger category based on semantic similarity, affording an opportunity for analysis which would not otherwise have presented itself. This example supports Rayson's (2008:543) argument in favour of categorising words into larger groups such as semantic domains, mentioned in 2.5.4.

However, the USAS tool did not successfully classify some of the meanings in the "Clothes and personal belongings" domain. For example, *pants* is a verb describing a behaviour, not related to clothing or belongings (line 9, Fig. 20), and there were similar examples of other verbs such as *cuff* (meaning to hit, not the band at the end of a sleeve). The word *heels* is a body part used in a metaphorical way, not associated with footwear (line 10, Fig. 20; line 4, Fig. 21). In line 7, Figure 21 *twill* is a contraction of *it wil*, masked by the absence of an apostrophe, which USAS therefore interprets as the noun referring to a type of fabric[61].

---

[60] The other two instances occur in the history play *Henry VIII* and the tragedy *Romeo and Juliet.*

[61] The *OED* indicates that the fabric *twill* is attested in the Early Modern period, in this and other variant spellings such as *tywlle*; see http://www.oed.com/view/Entry/208062?rskey=jFjMub&result=1&isAdvanced=false#eid (accessed 16.03.12).

The "Furniture and household fittings" domain is more reliable. Concordance extracts below provide examples from the *SDC* (Figure 22) and the *NDC* (Figure 23).

1.	half of a good play ! I am not	furnished	like a beggar , therefore to be
2.	beggar ; wouldst have made my	throne	A seat for baseness . No ; I ra
3.	wouldst have made my throne A	seat	for baseness . No ; I rather ad
4.	speak of him when he was less	furnished	than now he is with that which
5.	and the prime-roses Bear to my	closet	. Fare thee well , Pisanio : Th
6.	of door most rich ! If she be	furnished	with a mind so rare , She is al
7.	ble than that runagate to your	bed	, And will continue fast to you
8.	he leaf where I have left ; to	bed	: Take not away the taper , lea
9.	How bravely thou becom'st thy	bed	! fresh lily , And whiter than t
10.	have conquered my yet maiden	bed	, Remain there but an hour

Figure 22. Concordance extract for semantic domain H5 (Furniture and household fittings): Shakespeare's plays

1.	eeches . Shroud you within this	closet	, good my Lord , Some trick now
2.	now brother what travailing to	bed	to your kind wise ? I assure you
3.	I do commit you to your pitiful	pillow	Stuffed with horn-shavings . Bro
4.	g Alcumye . Thou shalt lie in a	bed	stuffed with turtles feathers ,
5.	u dishonour thus thy husband 's	bed	, Be thy life short as are the f
6.	r the soft down Of an insatiate	bed	. oh my Lord , The Drunkard afte
7.	ll his reverent wit Lies in his	wardrobe	, he 's a discrete fellow When h
8.	lting , he showed like a peuter	candlestick	fashioned like a man in armour ,
9.	on . May it thrive with you . A	Chair	there for his Lordship . Forbear
10.	women go to Church : Bear their	stools	with them . At your pleasure Sir

Figure 23. Concordance extract for semantic domain H5 (Furniture and household fittings): other contemporaneous plays

Some of the above examples are used in metaphorical contexts, e.g. the verb *furnished* (lines 1, 4 and 6, Fig. 22). The reference to a *throne* (line 2, Fig. 22) is a metaphorical reference to the reign of the monarch, as well as a literal reference to the monarch's chair or seat. In both corpora, *bed* is used literally (e.g. line 8, Fig. 22; line 4, Fig. 23), but in a metaphorical context to mean to have sex, and particularly sexual access to women (e.g. line 10, Fig. 22; line 6, Fig. 23).

The "Objects generally" domain is also reliable, and contains a wide range of nouns describing common everyday items used both literally and metaphorically in characters' dialogue (e.g. *rock*, *bell*, *knife*, *spoon*, *rope*, *cup*, *key*). This domain is notable amongst the others in Table 21 for showing different vocabulary in the two

corpora; in the others the vocabulary tends to be shared. The use of "household words" in Shakespeare's *Macbeth* is discussed by Hopkins, who argues that the appallingly homicidal nature of Macbeth and Lady Macbeth (a married couple) is dramatically offset by a parallel portrayal of "customary domesticity" (2004:259), through language surrounding meals and the hosting of their guests. This may well be true, but the locked domain results suggest that it is not necessarily unusual or an exclusively Shakespearean phenomenon. It is in fact difficult to imagine an audience being able to engage with a play which does not have, at some level, a thread of familiar activities which they can relate to in addition to the more extraordinary events of the plot.

The "Farming and horticulture" domain is quite heavily populated with literal references to concepts surrounding the pastoral themes of a few of the plays in both corpora (as Table 21 shows, *shepherd* is the most frequent word in this domain, in both corpora). This may account for the strength of the locking across the corpora, although other plays without pastoral themes also contain a great many references to horticultural or agricultural concepts. Examples are given in the concordance extracts from the *SDC* (Figure 24, below) and the *NDC* (Figure 25, on the next page).

1.	ic will work with him . I will	plant	you two , and let the fool make
1.	ady of the Strachy married the	yeoman	of the wardrobe . Fie on him ,
3.	Would not a pair of these have	bred	, sir ? Yes , being kept togeth
4.	when wit and youth is come to	harvest	, Your wife is like to reap a p
5.	harvest , Your wife is like to	reap	a proper man : There lies your
6.	towed upon me ; I saw't i' the	orchard	. Did she see thee the while ,
7.	e for him at the corner of the	orchard	like a bum-baily : so soon as e
8.	out to be of good capacity and	breeding	; his employment between his lo
9.	so excellently ignorant , will	breed	no terror in the youth : he wil
10.	his town , Where lie my maiden	weeds	: by whose gentle help I was pr

Figure 24. Concordance extract for semantic domain F4 (Farming and horticulture): Shakespeare's plays

1.	He spreads his bounty with a	sowing	hand , Like Kings , who many t
2.	Why here 's an end of all my	harvest	, he has given me nothing Cour
3.	n will . One were better be a	thresher	. Vdsedath , I would fain spea
4.	I love thee now ; if woman do	breed	man She ought to teach him man
5.	hs , when he questioned , may	breed	in him a jealousy , perchance
6.	called valiant , a word which	breeds	more quarrels than the sense c
7.	unto my house, I have an	Orchard	that hath store of plums
8.	nd sees , peradventure it may	breed	an offence to him . How can it
9.	off sir , she is not for your	mowing	. She is for your mocking . An
10.	will not . But what shall she	reap	hereby ? comfort in another wo

Figure 25. Concordance extract for semantic domain F4 (Farming and horticulture): other contemporaneous plays

Many farming and horticulture terms are used literally, to describe the surroundings of the play, e.g. *orchard* (lines 6 and 7, Fig. 24; line 7, Fig. 25). Other concepts are often used in metaphorical ways in the corpora, such as *harvest* (line 4, Fig. 24; line 2, Fig. 25), and *reap* (line 5, Fig. 24; line 10, Fig. 25). Table 21 at the start of this section shows that *bred* and *breed* are among the most frequent words in this semantic domain in the *SDC*, and that *breed* is also among the most frequent in the *NDC*. Examples of forms of the verb *breed* are shown in Figure 24 (lines 3, 8 and 9) and Figure 25 (lines 4, 5, 6 and 8). In well over than half the cases of *breed* in each corpus, it has meanings associated with producing or leading to a result of some kind (as in line 9, Fig. 24 and line 8, Fig. 25). However, in EModE drama, breeding also has sexual connotations as well as meanings associated with farming and horticulture. "Agriculture" is one of the concepts identified by Oncins-Martínez (2006, 2011) as a source domain for sexual metaphors in Shakespeare's plays and in wider EModE (from his 2006 dictionary data, mentioned in 3.3.3), so I examined the concordance data to see whether there was much evidence of this surrounding forms of the verb *breed* in my data. According to the *OED*[62], the meaning of *breed* in the Early Modern period could be associated with animate outcomes or non-animate outcomes, i.e. the results of sexual reproduction or

[62] See http://www.oed.com/view/Entry/23020?rskey=8wIzdl&result=2#eid (accessed 12.01.12).

non-sexual processes. In my data, non-animate outcomes are by far the most common, though I could not rule out the possibility that they are used in the context of sexual metaphors that are blurred by changes in meaning over time and/or a shift in ideologies about morality (difficulties pointed out in 3.3.3, based on Crystal and Crystal 2002, Oncins-Martínez 2006 and Tissari 2010a:138). Crystal and Crystal (2002:54-55) distinguish between meanings of growth, development and bringing into existence for the verb *breed* in Shakespeare's plays (as well as a fourth meaning of being brought up, examples of which are also present in my corpora). Onions (1983:23) does not classify the meaning of *breed* in the same way, however. He records a general meaning of keeping or supporting, which does not apply to most cases in my data, and a Shakespeare-specific use of "bred out", meaning "exhausted" or "degenerated", which is supported by my data since there are no examples of it at all in the *NDC*. That helps to account for the presence of *bred* as one of the most frequent words in the "Farming and horticulture" domain in the *SDC* but not the *NDC* (in Table 21).

Further investigation into concepts concerned with everyday life in EModE plays could usefully be carried out in future research. For example, the link between such concepts and the ways people address one another in Shakespeare's plays is made by B. Busse (2006). She argues that "[n]atural phenomena used vocatively make reference to animals, parts of the body, nature in general, and tangible objects, such as food, furniture, and clothing" (2006:332).

I now move on to some of the remaining locked domains in Table 21. The domain "Quantities: many, much" contains words which commonly have an evaluative function in characters' dialogue (both literal and metaphorical). These add to the evidence of Shakespeare and his contemporaries sharing preferences for frequently-

used kinds of hyperbolic language in dramatic dialogue (found in the locked clusters in 7.3). There are a few concepts that seem to be misclassified in this domain, such as *raging*. The "General and abstract terms" domain is solely populated by the words *thing*, *things* and *stuff* in both corpora, which are frequently-used generalisations, though some instances of *stuff* are verbs and therefore misclassified in this domain.

Some words and concepts in the domain "Unlikely", ranked 7th in Table 21, make a contribution to dramatic tension, through characters articulating doubt or uncertainty. The concordance data shows that characters in both sets of plays frequently discuss and evaluate what may or may not happen, and the likely consequences, using words such as *impossible*, *doubtful* and *uncertain*. However, the reliability of this domain is dubious, because the most frequently-occurring word in it in both corpora, *doubt*, often conveys the opposite meaning to uncertainty[63]. The concordance data shows that it is frequently used in a negated way to express certainty, e.g. in expressions such as *doubt not*, *no doubt, without doubt, I doubt not* and *I doubt it not*. These cases would be more accurately grouped with other concepts expressing certainty, such as *sure* (which would need to be done manually).

To sum up the locked semantic domains, the concepts in them which surround the weather and other areas of everyday life (clothes and belongings, furniture, farming and horticulture, and other objects) have an important function in literal descriptions of the surroundings of the plays and the people who inhabit them. However, they are also potentially rich sources of conceptual metaphors. Lakoff and Johnson argue that:

---

[63] *Doubt* is among the "false friends" identified by Crystal (2008:156-159), mentioned in 3.3.1, which he argues has a sense of meaning "fear, suspect" which applies in about 20% of all cases in which it is used as a verb in Shakespeare's plays.

> metaphors that are imaginative and creative [...] are capable of giving us a new understanding of our experience. Thus, they can give new meaning to our pasts, to our daily activity, and to what we know and believe. (1980:139)

Further insight into the ways dramatists of the late 16th and early 17th centuries accented their representation of social life could be gained from more detailed analyses of the kinds of concepts used in metaphorical contexts by Shakespeare and other dramatists of the same period. These would need to be supported by appropriate socio-historical evidence, however, to avoid misinterpreting historical metaphors on the basis of more familiar present-day ideas and assumptions.

While my results support the findings of Archer et al. (2009), Culpeper (2009) and Koller et al. (2008) with regard to the potential for identifying potential conceptual metaphors through semantic domain analysis, like them I also found that a great deal of further manual analysis, including the re-classification of many cases, would be required for the results to be reliable. As also documented by Archer et al. (2009) and Culpeper (2009), even with the benefit of the EModE tagger my findings show that the USAS tool has some trouble classifying the word meanings accurately, which reduces the reliability of some domains. Notwithstanding this, since I have focused on high-frequency results and directed my analyses to categories in which the majority of words are correctly classified, the locked domain results have provided some interesting and surprising insights into the play-texts, over and above the other types of locked results. They have revealed more than could be seen in the domains which are simply the most frequent when the corpora were examined independently of one another, in 6.3. The strength of references in both sets of plays to extreme dislike, and to the weather and other aspects of everyday social life did not emerge in those results.

I now summarise what the locked results in this chapter have shown about the style of Shakespeare, other contemporary dramatists, and the register of EModE

drama, and I draw some initial conclusions about what this particular method has shown about the compatibility of the corpora, and the value it adds to the study.

7.5 Discussion and conclusions

7.5.1 Implications for Shakespeare's style and for the register of Early Modern English drama

Although it was reasonable to anticipate that there would be similarities in the language in both corpora, since they contain texts from the same genre and period (albeit by different authors), the analyses of locked results in 7.2, 7.3 and 7.4 have furnished some specific details about the features which are similar. Some of them reflect the "aboutness" of the play-texts (as is also typical in key results; see 2.7), such as the lockwords MAJESTY, POWER and DEATH, in 7.2. They are not localised, though, since they are topics and themes for which Shakespeare and other dramatists shared a preference, in all three genres, and are therefore associated with the register of EModE drama. The lockword FAREWELL (7.2), is also evidently a register feature, through having a clear function of stagecraft (setting up the departure of characters from the stage, so the plot can move on). These results are perhaps not surprising, given what is already known about the popular themes in EModE drama and about the pragmatics of drama. However, other locked results were more remarkable, notably the strong shared preference between Shakespeare and other contemporaneous dramatists for:

- a non-deferential term of reference and address (the lockword FELLOW, in 7.2), contrasting with the frequently-used deferential terms which were seen in 6.2, such as SIR and LORD;

- a certainty marker (the locked 3-word cluster I AM SURE), since certainty did not feature among the most highly-frequent Modalising clusters in 6.3 (and indeed most of the highly-frequent clusters were topical); and
- concepts of ordinary everyday life in the semantic domain results, such as the weather, clothing, furniture, household objects and farming (in 7.4), again none of which occur among those which are simply the most frequently-occurring in 6.4.

The locked semantic domains were particularly rewarding, grouping concepts for which Shakespeare and other contemporaneous dramatists shared a preference, and thereby highlighting some areas which have not been addressed in much detail in other studies but which now seem worthy of closer attention. The weather, for example, is one of the most strongly locked domains (in 7.4), and although weather has been noted as being of interest by other scholars (e.g. Richmond 2002), the extent to which it is harnessed in figurative language in the register of EModE drama has perhaps been under-appreciated. In particular, a study of the personification of weather, in the form of pathetic fallacy, would be of interest in showing how Shakespeare and other dramatists use it to create vivid images for the audience (which I suggested in 7.4 could be considered linguistic "special effects"). Similarly, although the concept of love has been investigated in Shakespeare's plays using key domain analysis (by Archer et al. 2009), the locked domains indicate that a study of its emotional antithesis, extreme dislike, would also be worthwhile, since it is also one of the most strongly locked conceptual areas in both sets of plays, with a function of adding to the sense of drama through overlaying the topics and themes of the plays with emotion, and in revealing characters' motivations (argued in 7.4).

Through the locked results, it has been possible to make some brief links with studies in other research areas:

- historical sociolinguistics (the lockword FAREWELL and Arnovick's 1999 research, in 7.2); and
- literary critical studies (e.g. the locked semantic domain of clothing, and discussions of dress and costume in Jardine 1983 and Orgel 1996), in 7.4.

With more space, more and deeper analyses could have been carried out. Those I have included in this chapter illustrate a variety of ways in which locked results can add value to stylistic analysis, and potentially beyond. The analyses of the locked results add a whole new dimension to the investigation of Shakespeare's language style, by giving an empirically-based perspective of the extent to which it is similar to that of an aggregated group of his contemporaries.

7.5.2 Corpus compatibility issues highlighted by the analyses of locked results

The nature of the locking concept means that compatibility problems are much less likely to be visible through locked results than through key results (the subject of the next chapter). This is because locked results are oriented to similarity, and are those which match the most in both corpora, whereas key results are oriented to difference, and are those that match the most in one corpus but not in the other. A statistical method which is oriented to difference is more likely to turn up differences arising from problems such as textual incompatibility, as well as those that arise from stylistic variation. Essentially, because the locking method is not being asked to find differences between the corpora, it is predisposed to finding language in the plays which is unaffected by compatibility problems. This must be borne in mind in a study of historical texts where there are known issues of spelling standardisation (see 5.4)

and/or textual compatibility (see 4.2.2). Few problems surface with the locked results, apart from some categorisation issues noted in the semantic domains in 7.4 (such as the word *rime*). It is difficult to say whether or not the locked results would have been different in any way if the punctuation and spelling were entirely compatible across both corpora. However, as discussed in 5.4, substantial effort was put into making the texts as compatible as possible. This remains a limitation of the study but, as I have shown in this chapter, the locked results that were identified are certainly of interest and use for stylistic analysis.

7.5.3 Evaluating the locking method

The investigation of locked results in just two corpora has been straightforward to carry out, requiring simple adjustments to the parameters of two software tools, *WordSmith* and *Wmatrix*, which compute key results based on the log-likelihood statistical test. Comparisons of similarities in more than two corpora would require more effort on the part of the user, for example by using Baker's (2011:72) method with *SPSS*, based on calculations of standard deviation and co-efficient of variance, or Rayson's spreadsheet which compares log-likelihood scores[64]. However, the keyness tools offer the advantage of automatically compensating for different file sizes. Through a direct statistical comparison, the lockwords, locked word clusters and locked semantic domains have yielded much more information about similarities between Shakespeare's style and that of other contemporaneous dramatists, with corresponding implications for the register of EModE drama, than was obtained from the frequency counts in chapter 6 (when the two corpora were examined independently).

[64] See http://ucrel.lancs.ac.uk/llwizard.html) (accessed 14.03.12).

Despite being the statistical opposite of keywords, the lockwords fall into similar varieties as those usually found in keywords output (see Scott 2000 and my discussions in 2.7). The "aboutness" results are thematic, rather than topical, since they are not localised to one or a few plays. There is one proper noun among the lockwords (JOHN, in 7.2), as well as a number of other results which afford profitable stylistic analysis. Some of the results support intuitions about the register of EModE drama, for example the recurrent themes of power, death and noble or royal characters (7.2). Others have been surprising, notably that *fellow* is the single word with the most similar frequency, statistically, across both corpora (7.2), and that the semantic domains with the most similar frequency contain concepts surrounding extreme emotional dislike and the weather (7.4).

The locked results have beneficially highlighted some similar language features in the corpora which I would argue could not have been anticipated, for the following reasons:

- They seem not to be prominent in the play-texts (i.e. they do not stand out psychologically, in Leech and Short's 2007 terms, although as noted in 1.2.2 a contemporaneous audience or reader of the plays would doubtless have had greater insight).
- This means, therefore, that they are similarities based on patterns which are too complex to be discovered through manual effort, and which can only be viewed with the help of automated counting processes using large amounts of dialogue.
- These similarities do not, however, surface through a comparison of computer-generated frequencies when the corpora are examined independently of one another; they only emerge when the log-likelihood statistical test is applied to

the corpora together, and the corpus tools adjusted to identify the items with the most similar frequency, statistically.

The reasoning above supports the case for using locking method for stylistic investigation, since the results have led to analyses of words and concepts in EModE drama which have not been the focus of other studies. These clearly have implications for the register of EModE drama, as summarised in 7.5.1.

Through providing specific, empirically-based details about language which deviate the least, statistically, between the two corpora, the locked results arguably suggest language norms between Shakespeare's plays and plays by other contemporaneous dramatists. The definition of what constitutes a "norm" is not of course fixed, and may vary with different methods and statistical measures. However, language which is the least statistically deviant is a useful comparator for that which deviates the most, such as is identified through keyness analysis (the focus of the next chapter). This is relevant to theories of foregrounding in stylistics, as I argued in 1.1 and 2.6.1. Foregrounding is a "relational" concept (van Peer 1986:7; see also Mukařovský 1964a and b), and while the keyness method can assist the investigation of foregrounded language (through identifying language features which deviate, statistically, between texts or corpora), the actual norms remain unknown. Keyness and locking are opposite concepts, statistically (Baker 2011:73; see 2.6.2), and by virtue of this the locking method provides some context for foregrounding, by identifying language which is arguably backgrounded through statistical non-deviance.

The analyses in this chapter also confirm that Baker's (2011) locking concept can be applied successfully to synchronic corpora as well as to diachronic corpora, and that the concept and method can be extended to recurrent word combinations and to words which group into semantic categories (as well as to single lockwords). The

outcomes of the analyses in 7.2-7.4 demonstrate that there is much to be learned from examining similarities as well as differences between corpora. This supports Baker's (2004:349) point that keyness offers essentially a one-sided view of language in corpora in orienting to difference only. As argued in 2.6, this point has largely not been taken up in corpus stylistic research. An examination of similarities is essential, in order to reach a deeper and clearer understanding of the implications of the differences, which are the focus of the next chapter.

CHAPTER 8. INVESTIGATING DIFFERENCES BETWEEN SHAKESPEARE'S PLAYS AND OTHER CONTEMPORANEOUS PLAYS USING KEYWORDS, KEY WORD CLUSTERS AND KEY SEMANTIC DOMAINS

8.1 Introduction

In this chapter, I examine the opposite and complementary phenomenon to the locked results in chapter 7, by investigating key results (the method of which was explained in 2.5). Through these, I now explore the ways in which Shakespeare seems to part company with his contemporaries in his choice of language for dramatic dialogue, having identified some shared language preferences in the previous chapter (based on statistical frequency and similarity). Again, the layout and principles of this chapter follow those set out in 6.1. I discuss keywords in 8.2, key 3-word clusters in 8.3, and finally key semantic domains in 8.4. A concluding discussion follows in 8.5, which includes an overall assessment of the method of semantic domains, based on the outcomes of this and the previous two chapters.

8.2 Keywords in Shakespeare's plays

The most statistically significant (key) words in the *SDC* when it is compared to the *NDC* are shown in Table 22 on the next page, with the raw frequencies for both corpora. The results are ranked in descending order of keyness (log-likelihood, or "LL") value. Positive key words (those which occur in Shakespeare's plays relatively more often than would be expected) are shown in the left-hand column, and negative key words (those which occur relatively less often than would be expected) are shown in the right-hand column.

Table 22. Top 20 rank-ordered keywords in Shakespeare's plays (minimum frequency=200; p=0.000001)

Rank	Positive key words	*SDC*	*NDC*	LL	Negative key words	*SDC*	*NDC*	LL
1	TIS	1,386	49	1,624.5	WHO	1,133	3,502	1,168.1
2	AY	770	28	898.1	YE	277	967	374.9
3	CÆSAR	346	0	495.6	SEE	1,416	2,303	175.7
4	O	2,561	1,591	275.5	UNTO	422	848	127.1
5	ANTONY	205	4	259.4	HA	221	489	92
6	YOURSELF	276	29	242.6	CAN	1,148	1,724	91.8
7	TOMORROW	218	19	205.4	YES	207	459	86.7
8	ITSELF	240	27	205.1	NOW	2,739	3,607	83.3
9	YORK	216	20	199.2	THEN	2,096	2,816	76
10	THAN	1,834	1,173	178.3	MISTRESS	392	680	65.9
11	THE	26,316	24,511	177.5	THESE	1,297	1,782	56.4
12	FATHER'S	232	48	140.1	I'LL	1,735	2,300	56
13	SPEAK	1,157	727	119.6	COURT	243	451	54.3
14	HIM	5,023	4,211	113.7	OR	2,305	2,931	49.5
15	HE	6,184	5,350	104.8	ALL	3,587	4,386	48.3
16	WHICH	2,177	1,646	100.3	FIRST	527	809	47.9
17	MOST	1,154	763	99.1	MAY	1,607	2,111	47.7
18	VERY	807	481	98.9	ONLY	306	509	42.3
19	DOES	324	131	93.5	LONG	456	699	41.1
20	BESEECH	221	75	82	FAITH	392	617	40.9

A few of the keywords in Table 22 arise from variation in spelling and punctuation between the two corpora, and would not lead to reliable conclusions regarding the language styles of the playwrights. I mention these briefly, before going on to results of more stylistic interest. As explained in 5.4, punctuation, in particular, could not be standardised very far between the texts in both corpora using automated methods. Also, as discussed in 4.2.2, the *SDC* texts are based on a 1916 edition of the plays which has undergone some modernisation, through which the spelling and punctuation has been standardised further than that in the early extant *NDC* texts.

In Table 22, the positive keyword TIS is standardised to one form in the *SDC* but split over several variant forms in the *NDC* (both with and without an initial apostrophe), so its apparent over-representation in Shakespearean plays cannot be relied upon. This is also the case with the positive keyword FATHER'S, because of the

standardisation of s-genitives with an apostrophe in the *SDC* but not in the *NDC*. The negative keyword I'LL is reliable, since the spelling variant *Ile* was regularised to *I'll* or *isle* as appropriate in the *NDC*, and it is already a standardised form in the *SDC*, so *WordSmith* is therefore successfully matching all the instances in both corpora.

There are reliability issues with AY and O, which are positive keywords in Table 22, and with HA, a negative keyword. Multiple variants of AY and O in the *NDC* (*ay*/*aye* and *O*/*Oh*) cause them to be apparently over-represented in the *SDC*. This could be remedied for future research in the case of AY, because Culpeper and Kytö (2010:222) find both forms in the *CED* and argue that they could be regularised to one spelling. However, in 6.2 I noted that the variant forms *O* and *Oh* were not regularised to one spelling in the *NDC* because they do not entirely overlap in where and how they are used (based on Culpeper and Kytö's 2010:278 detailed examination of the *CED*). I cannot be certain whether any cases of *O* or *Oh* would have been standardised by the compositors of the SDC texts, blurring distinctions that might have been present in early extant editions. Although *ha* appears to be under-used in Shakespearean drama, the concordance data for the *NDC* shows that this form not only functions as a pragmatic marker, but as a contracted form of *have*, with or without an apostrophe after it. There are many more instances of this contraction in the *NDC* than the *SDC*, which account for the negative keyness, but it is possible that this contraction has been standardised out of the *SDC* texts during the modernisation process.

The apparent over-representation of TOMORROW, YOURSELF and ITSELF in the *SDC* is affected by non-standard compound forms in the *NDC* (e.g. *to morrow*; see 5.4 and 6.2). Finally, the most negative keyword in Table 22, WHO, is unreliable because its over-representation in the *NDC* is due to the fact that *WordSmith* unaccountably picks up the word *who* from the XML speaker-id tags (see 3.3.2).

The incompatibility of punctuation such as the apostrophes causes minimal inconvenience to the study, as indicated in 5.4, since they arise in so few of the results. Those results with possessive apostrophes can be excluded from my analyses without any major consequences. However, the exclusion of some of the other potentially unreliable results in Table 22 is more of a loss. For example, if they are genuinely key in Shakespeare's dialogue, TIS and AY would be worth investigating as possible markers of dialect and/or social rank. For example, Culpeper (2009:37) finds *tis* to be a keyword in the dialogue of Capulet, when all his speech is compared to that of five other characters in Shakespeare's *Romeo and Juliet* (see 2.5.2). An investigation of the pragmatic markers *O* and *ha* would also be of interest. It would extend Culpeper and Kytö's (2010) findings from EModE drama comedy in the *CED* to the history and tragedy genres too, and Culpeper's (2009) finding that *O* has expressive functions which contribute to characterisation in *Romeo and Juliet*. As argued in 5.4, however, the process of standardising texts from this historical period is inevitably imperfect, and remains a limitation of applying computer-based methods to them.

Having acknowledged the issues surrounding a minority of unreliable keyword results, it is important to emphasise that the majority of results in Table 22 are reliable, and not affected by compatibility issues between the corpora. My investigations into the concordance data confirm that they are linguistic constructs which are over-represented (positive keywords) or under-represented (negative keywords) in Shakespeare's plays compared to the other contemporaneous plays. They are not all potentially stylistically interesting, however. Despite setting the corpus linguistic software parameters to minimise topical or localised results (see 3.4), there are some proper nouns and "aboutness" results which arise from the topics of the plays, as is typical of key results output (discussed in 2.7). ANTONY and CÆSAR are the names

of two major protagonists in the Shakespeare corpus. YORK is the name of a place and a noble family which features in multiple plays in the *SDC*. The rival "houses" of Lancaster and York are central to over half Shakespeare's history plays. Other dramatists do feature the York/Lancaster storyline (e.g. Heywood's *Edward IV Part I*, which is in the *NDC*), but not nearly to the same extent. The under-representation of COURT in Shakespearean drama is likely to be due to the locations or settings of the plays, although court settings feature in plays in both corpora. These results are not really useful for my analysis of language styles. They could have been further minimised by manually excluding any results which occur in, say, fewer than five different plays. This would allow more of the stylistically interesting key results to surface nearer the top of the list, and would be a step worth taking in more detailed analyses of language styles in smaller component sections of the corpora. I now come to the results in Table 22 which are potentially stylistically interesting.

The over-representation of BESEECH, a deferential verb occurring as a positive keyword in the Shakespearean dialogue, ties in with Craig and Kinney's (2009b:38) finding that Shakespeare uses *beseech* relatively more than other contemporaneous dramatists. It is therefore worth investigating further. BESEECH is used in requests made by characters of varying social rank, gender, and in different kinds of situations, some of which are given in the concordance data in Figure 26.

N	Concordance	File
1	onsent goes not that way. I beseech you heartily, some of you go home	\scmww.txt
2	t she. For what reason, I beseech you? For this re	\scshrew.txt
3	ning with me When you are by at night. I do beseech you- Chiefly that I might set it in	\sctemp.txt
4	Welcome, dear Proteus! Mistress, I beseech you, Confirm his welcome with s	ctwogen.txt
5	l all that speak of it. I do beseech you, madam, be content.	ar~1\shkj.txt
6	become the grave. I do beseech your majesty, impute his words T	1\shrichii.txt
7	me? I am unfit for state and majesty: I do beseech you, take it not amiss, I cannot no	shrich~1.txt
8	ve, what trade? Nay, I beseech you, sir, be not out with me: yet, if	ar~1\stjc.txt
9	stray To match you where I hate; therefore, beseech you To avert your liking a more w	~1\stlear.txt
10	I do beseech you, sir, troubl do beseech you, sir, trouble yourself no further.	r~1\stoth.txt

Figure 26. Concordance extract for *beseech*: Shakespeare's plays

Figure 26 show BESEECH being used to mitigate requests for information from an addressee (line 2); in requests to modify behaviour or attitude (e.g. to be content, line 5); and in requests for actions (e.g. to leave, line 1). A dispersion plot for BESEECH in the *SDC* is given in Figure 27 below. This indicates that although BESEECH is distributed fairly widely across the three dramatic genres, it occurs in tighter clusters in some of the less prototypical comedy plays between about a quarter and halfway through them (*Cymbeline*, *Measure for Measure* and *The Winter's Tale*). It also clusters near the beginning of the history play *Henry VI Part II*, but nearer the end in *Henry IV Part II* and *Henry V*. There is also a cluster near the start of the tragedy *Antony and Cleopatra*. (See Appendix IV for a full list of text-ids for the plays).

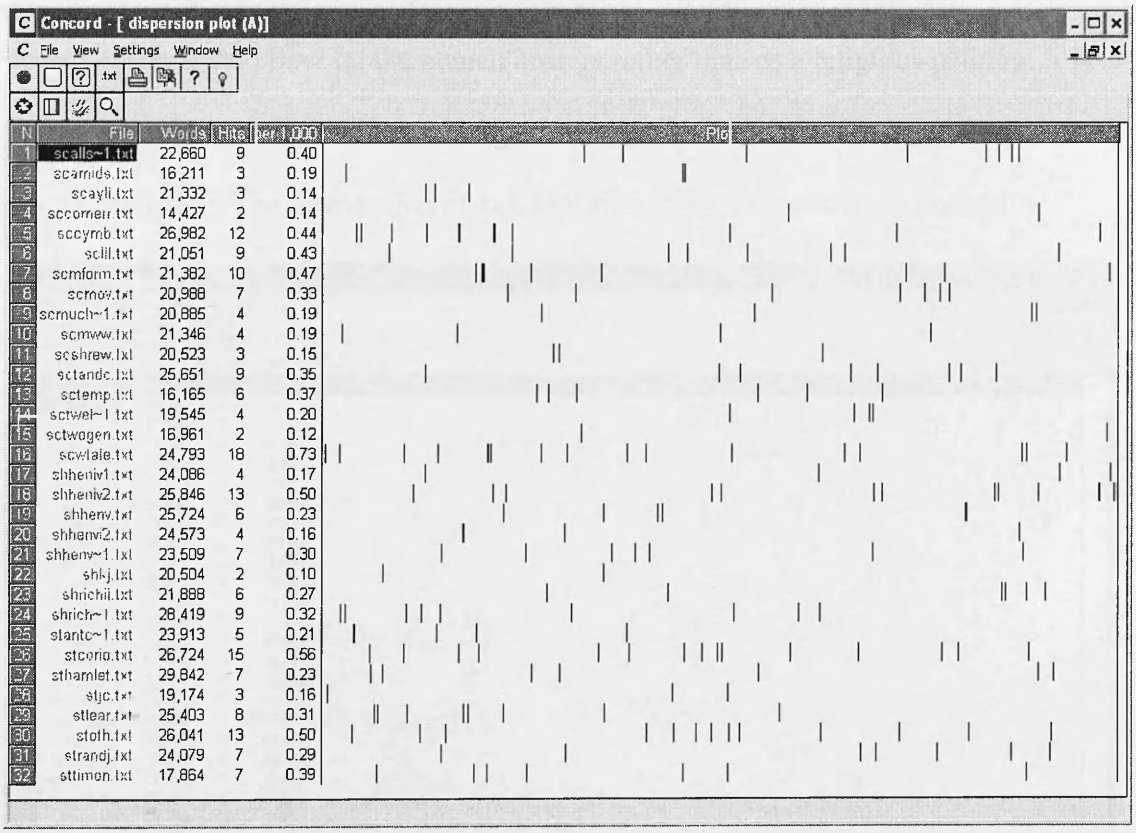


Figure 27. Dispersion plot for *beseech* in the *SDC*

Culpeper and Archer (2008:72, citing Fraser 1975:197) argue that *beseech* acknowledges the requester's position to be relatively powerless compared to that of the addressee. In my Shakespearean data, *beseech* does not necessarily occur in

situations where the speaker is of lower social rank or in a position of lower power than the addressee(s), though that is sometimes the case. However, it also seems to be used for an expressive effect, in situations where speakers have equal social rank and power. This is shown in example (27), from *Antony and Cleopatra.*

(27) Charmian: O! let him marry a woman that cannot go, sweet Isis, I beseech thee; and let her die too, and give him a worse; and let worse follow worse, till the worst of all follow him laughing to his grave, fifty-fold a cuckold! Good Isis, hear me this prayer, though thou deny me a matter of more weight; good Isis, I beseech thee!

Shakespeare, *Antony and Cleopatra*, I:ii (*SDC*)

In (27), Cleopatra's maid Charmian uses *beseech* in an appeal to the goddess Isis. While it would not be surprising for a deity to be addressed humbly using *beseech*, this speech is made for effect on the human hearers rather than as a religious petition. It is part of a frank, light-hearted discussion of marriage and fidelity with another maid and a male attendant. The distribution of *beseech* in the *NDC* is shown in Figure 28.

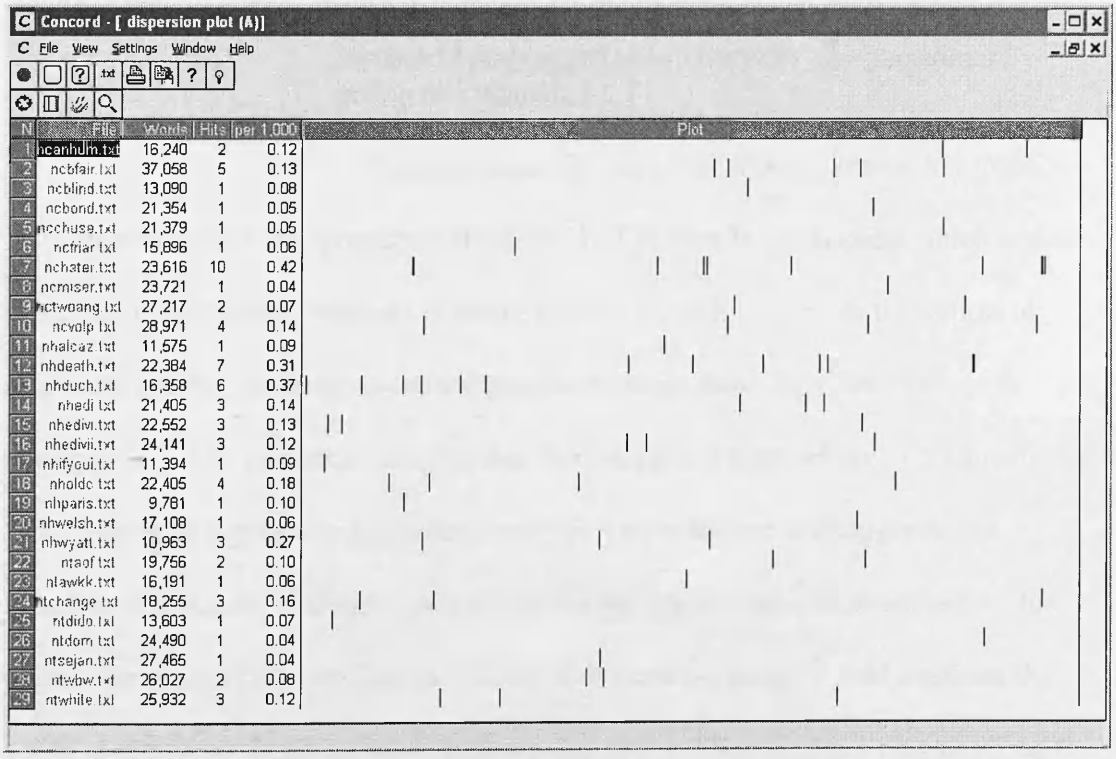


Figure 28. Dispersion plot for *beseech* in the *NDC*

Figure 28 shows a much lighter distribution, as would be expected from the keyness of BESEECH in the *SDC*, but again it is not concentrated particularly in any one genre. However, in contrast to the *SDC*, the only visible concentrations of BESEECH in the *NDC* are near the start and end of a comedy: Beaumont and Fletcher's *The Woman Hater*. As shown in example (28) below, from near the end of the play, it is used in succession by two characters of lower social rank (Gondarino, a male, and Oriana, a female) to an aristocrat (the Duke of Milan), in a situation of gravity and serious consequence. The Duke is in a position of power to determine what punishment will be meted out to Gondarino for spreading false rumours about Oriana.

(28) Oriana: I do beseech your Lordship, for the wrongs this man hath done me, let me pronounce his punishment.
Duke of Milan: Lady, I give to you, he is your own.
Gondarino: I do beseech your grace, let me be banished with all the speed that may be.
Oriana: [...]
Lord Gondarino, you have wroug'd me highly, yet since it sprung from no peculiar hate to me, but from a general dislike unto all women, you shall thus suffer for it;
Gondarino: [...]
My Lord I do beseech your Grace for any punishment saving this woman, [...]

Beaumont and Fletcher, *The Woman Hater*, V:i (*NDC*)

It is interesting that the concentration of BESEECH here is in a request which has life-or-death consequences, whereas in Shakespeare's plays it clusters in the course of much less serious requests, or indeed pseudo-requests made for effect such as in example (27). My evidence indicates that that *beseech* is used relatively frequently by Shakespeare in a generally hyperbolic way, as a style feature to exaggerate the humility of requests, whereas it is reserved for particularly serious situations by his peers. The sense of desperation associated with graver requests would augment the sense of drama surrounding more ordinary requests in Shakespeare's plays, raising the emotional level. I mention *beseech* again in the next section, since it occurs in several

key clusters; it is clearly linked to the greater number of most frequently-occurring deferential 3-word clusters noted in Shakespeare's plays in 6.3.

MISTRESS is also a deferential term, but it occurs relatively less frequently in Shakespeare's plays (as a negative keyword, Table 22), despite the fact that there are similar numbers of female characters in both corpora. This contrasts with an over-representation of HE and HIM, both of which are positive keywords in the Shakespearean drama. This combination of evidence suggests, at face value, that women are referred to relatively less in Shakespeare's plays than in the other contemporaneous plays. However, there are other possible contributory factors. It might simply be the case that Shakespeare uses more proper nouns instead of the pronouns and vocatives that occur as key in my data. Also, the social rank of the women in the corpora has a potential effect on the relative frequency of use of MISTRESS, since it is a courtesy title not a pronoun. If there are more lower-ranking characters in the *NDC* who would conventionally address higher ranking women using a deferential term, or indeed more higher-ranking women who merit deferential address, that could account for the keyness. Unfortunately the resources of the project did not allow for social rank to be included in the annotation process (discussed in 5.2). This would need to be considered in any detailed investigations of language associated with deference, however. Furthermore, as noted in 4.3.2.2, to fulfil the criteria for balancing the amount of female dialogue in the *NDC* with that in the *SDC*, some of the other contemporaneous plays that are included have particularly female-oriented topics, for example *The Woman Hater*, *A Woman Killed With Kindness* and *Women Beware Women*. Therefore, relatively more talk about women in the other contemporaneous plays could reflect a topical or "aboutness" contrast between the corpora contents, rather than a language style feature.

The negative keyness of SEE, in Table 22, confirms the indications in 6.2 that it is used relatively less in Shakespeare's plays than in other contemporaneous plays (although it is very frequent in both corpora). As discussed in 6.2, characters often make claims on the basis of what they "see" (which is sometimes a metaphorical substitute for "know", also noted in 6.2 as being very frequent). A character voicing what he or she sees can be a useful dramatic device which explains information the audience needs to know to follow the plot and understand the characters, although it is one which appears to have been less favoured by Shakespeare than by his peers.

The concordance data shows that the negative keyword FAITH is used in similar ways in both corpora. It reinforces what the speaker is saying (e.g. in *i'faith*, *in faith*, *by my faith*). Crystal and Crystal (2002:435) list it as a widely-used form of swearing in Shakespeare's plays, so it is notable that it nevertheless appears to have been used relatively less often in dialogue by Shakespeare than by other dramatists. Interestingly, Culpeper (2009:37, 50-51) finds that *faith* is a keyword in the dialogue of the Nurse in Shakespeare's *Romeo and Juliet*, a character of relatively low social rank who, Culpeper argues, has "a tendency to dramatize events she narrates" (2009:51). A more detailed analysis of the kinds of speakers who use *faith* (e.g. their social rank and gender) would therefore be necessary to draw firmer conclusions about its relatively low frequency in the *SDC*.

The keyword results have usefully highlighted some compatibility issues between the *SDC* and the *NDC*. There are also a few indicators of Shakespeare's authorial style among the keywords, most firmly in the relative over-use of *beseech* (used to exaggerate the deference in relatively ordinary requests) and the relative under-use of *see*. Next, I examine key 3-word clusters in Shakespeare's plays.

8.3 Key 3-word clusters in Shakespeare's plays

Table 23 shows the top 20 key 3-word cluster results when the *SDC* is compared to the *NDC*, with raw frequencies. Again, they are ranked in descending order of keyness (log-likelihood, or "LL") value, with positive and negative results listed separately. (There are only four negative key results.)

Table 23. Top 20 rank-ordered key 3-word clusters in Shakespeare's plays (minimum frequency=50; p=0.01)

Rank	Positive key clusters	*SDC*	*NDC*	LL	Negative key clusters	*SDC*	*NDC*	LL
1	DUKE OF YORK	41	2	45.2	IT IS BUT	16	67	31.4
2	FARE THEE WELL	40	3	39.5	IT MAY BE	35	97	27.6
3	SO PLEASE YOU	26	0	37.2	THE KING OF	29	73	17.7
4	I PRAY YOU	242	141	31.6	AND IT IS	23	60	15.5
5	TIS NOT SO	21	0	30.1				
6	SIR JOHN FALSTAFF	19	0	27.2				
7	THE ISSUE OF	19	0	27.2				
8	TIS NO MATTER	19	0	27.2				
9	I HAVE SPOKE	19	0	27.2				
10	THE MARKET-PLACE	19	0	27.2				
11	FOR THE WHICH	19	0	27.2				
12	AIN'T PLEASE YOUR	18	0	25.8				
13	FARE YOU WELL	74	27	24.9				
14	I BESEECH YOU	74	27	24.9				
15	MY GOOD LORD	131	66	24.8				
16	DO BESEECH YOU	33	5	24.3				
17	DUKE OF GLOUCESTER	17	0	24.3				
18	WHAT IS THE	58	18	24.0				
19	THE DUKE OF	111	53	23.7				
20	I DO BESEECH	54	16	23.5				

As with the keywords in the previous section, some results in Table 23 reflect non-stylistic issues. Since TIS is an unreliable keyword due to variant forms in the *NDC* (see 8.2), the positive key clusters TIS NOT SO and TIS NO MATTER are also potentially unreliable. The reliability of the positive key clusters THE MARKET-PLACE and AIN'T PLEASE YOUR are also uncertain, because of the standardised punctuation to modern forms in the *SDC*.

I show the reliable results from Table 23 according to function, in Table 24 (in the same manner as for high-frequency and locked 3-word cluster results in the previous chapters). Negative key results are indicated by a minus sign in brackets after the raw frequency; positive key results are unmarked.

Table 24. Functions of top 20 key 3-word clusters in Shakespeare's plays

Interpersonal	Speech act-related	*Directive* FARE THEE WELL (40) I PRAY YOU (242) I BESEECH YOU (74) FARE YOU WELL (74) DO BESEECH YOU (33) I DO BESEECH (54) *Vocative* MY GOOD LORD (131)
	Modalising	*Downtoners/amplifiers/hedges/emphatics* SO PLEASE YOU (26) IT IS BUT (16-) IT MAY BE (35-)
Textual	Discoursal	*Question* WHAT IS THE (58)
	Narrative-related	*Reporting/reported clause fragments* I HAVE SPOKE (19)
	Organisational	*Informational elaboration* FOR THE WHICH (19)
Ideational	Topical	*People* DUKE OF YORK (41) SIR JOHN FALSTAFF (19) THE DUKE OF (111) DUKE OF GLOUCESTER (17) THE KING OF (29-) *Informational specificity* THE ISSUE OF (19) *States* AND IT IS (23-)

There is a modest presence of topical results in Table 24, some of which are proper nouns, but on the whole the constraints imposed to minimise them (see the discussion of corpus tools settings in 3.4) have worked. They are concentrated in the Ideational: Topical: People category, and are the titles or names of actual characters who occur in multiple Shakespearean plays (DUKE OF YORK, DUKE OF GLOUCESTER and SIR JOHN FALSTAFF), or fragments of a frequently-occurring rank of Shakespearean character (THE DUKE OF). When combined with the negative keyness of THE KING OF, these results suggest that dukes are over-represented in the *SDC*, while kings are under-represented, in comparison to the particular collection of plays comprising the *NDC*. Nevertheless, the results show that both sets of plays often centre around the activities of royalty and nobility (i.e. there is a similar axis of aboutness). With regard to other types of clusters arising from the topics of the plays, it is interesting to note how few Ideational: Topical: States clusters occur as key, now the corpora are being compared with one another statistically. Many States clusters occurred when the most frequent clusters were compared side by side in lists generated independently from each corpus (Table 15, 6.3), but these now cancel one another out in the keyness computations. This indicates that there is a common core of States clusters which are versatile formulae for building dramatic dialogue, used by other contemporaneous dramatists as well as by Shakespeare.

There is further evidence in the key clusters to support the idea that the language in Shakespeare's plays is relatively more deferential than that in the other contemporaneous plays. This idea was mooted in 6.3 with regard to the greater presence of deferential 3-word clusters in Shakespeare's plays, and strengthened in 8.2 by the keyness of BESEECH in Shakespeare's plays. I now pick up this analytical thread once more, using the key clusters in my data which contain the verbs *beseech*

and *pray*. The language formulae *I beseech you* and *I pray you* have been highlighted in other studies of EModE speech-related texts as being associated with the speech act of requesting, and are discussed, for example, among other parenthetical forms in Włodarczyk's (2007:121-127) EModE courtroom trial data. Włodarczyk (2007:126) reports findings which support earlier arguments of Brown and Gilman (1989:183) and Culpeper and Kytö (2000:55) that *I beseech you* is more formal and more deferential than *I pray you*.

Table 24 shows that my data features several Interpersonal: Speech-act-related: Directive clusters which contain the verb *beseech* (I BESEECH YOU, DO BESEECH YOU and I DO BESEECH; the latter two overlap to form the longer formula *I do beseech you*). Some examples of *beseech* clusters used in requests in Shakespeare's plays were shown in the concordance extract in Figure 26 in the previous section, and the extract from Charmian's speech in *Antony and Cleopatra* (in example 27) shows the cluster I DO BESEECH in use. I will not therefore present further examples of *beseech* clusters here, but will briefly report breakdowns of the results which I carried out to see whether beseeching appears to be located in the dialogue of any particular group(s) of Shakespearean characters.

My data shows that the clusters are key in the male Shakespearean dialogue (when it is compared to female Shakespearean dialogue as a reference corpus, and also when it is compared to male dialogue in the other contemporaneous plays). When gender comparisons are carried out in each dramatic genre, one or more *beseech* clusters occur in comedies, tragedies and histories. Therefore, beseeching is a language style feature of male characters in Shakespeare's plays, but not associated particularly with genre. It would be worth analysing the *beseech* clusters according to the social rank of speakers and addressees, in future research.

A preference for using *I pray you* in dramatic dialogue was highlighted as a potential Shakespearean style feature by the high-frequency 3-word clusters in 6.3, supporting the similar findings of Culpeper (2011:73). The verb *pray* underwent grammaticalisation during the late Middle English and Early Modern periods (Akimoto 2000:68; Traugott and Dasher 2002). Forms of *pray* used as a pragmatic marker are found in EModE drama (Akimoto 2000; T. Walker 2007:270-278) and, as indicated above, other speech-related text-types such as courtroom trial data (Kryk-Kastovsky 2000:215; Włodarczyk 2007). *I pray you* is one of the two most complex *pray* forms (the other being *I pray thee*). Less complex forms include *I pray*, *pray you* and *pray thee*, and the simplest forms are *pray* or *prithee*, used on their own (Lutzky and Demmen, forthcoming). Culpeper and Archer (2008:74-76) argue that *pray* forms are used as "support moves" in requests in EModE, and that the complex form *I pray you* conveys the sense of "an act of supplication" (2008:76; see also Jucker 2002:224). Since more than half the requests in Culpeper and Archer's data do not contain a support move of any kind, however, they argue that the mitigation of requests was not automatically a social requirement, particularly among speakers of higher social rank (2008:74)[65]. The reasons why Shakespeare chose to include a particularly humbling *pray* form relatively frequently in the dialogue of his characters are therefore worth investigating further.

I begin with concordance extracts (on the next page) providing some examples of I PRAY YOU in Shakespeare's plays (Figure 29) and the other contemporaneous plays (Figure 30), which illustrate its use in the context of making requests. I follow Culpeper and Archer (2008:45; 47-48) in considering a "request" to be one of Searle's (1969) "directives", which includes "requests" and "commands".

[65] The background to *pray* was researched by myself and Ursula Lutzky (Birmingham City University) between 2010-2012 using Lutzky's diachronic corpus of EModE drama comedy samples, not the corpora in the present study. See Lutzky and Demmen (forthcoming).

N	Concordance		File
1	n, hoping to be the wiser	by your answer. I pray you, sir, are you a courtier?	1\scalls~1.txt
2	ompany?	No epilogue, I pray you; for your play needs no excuse.	1\scamids.txt
3	you; you shall have some part of your will:	I pray you, leave me. I	r~1\scayli.txt
4	ate fashion.	My lord, I pray you, hear me.	\shheniv1.txt
5		Who may that be, I pray you? Thom	\shhenv~1.txt
6	What was your dream, my lord?	I pray you, tell me.	\shrich~1.txt
7	not drink.	I will, my lord; I pray you, pardon me.	1\sthamlet.txt
8	[Knocking within.] Anon, anon!	I pray you, remember the porter.	1\stmacb.txt
9	sir, how?	Are these, I pray you, wind-instruments?	ar~1\stoth.txt
10	e that should be husband comes to woo.	I pray you, tell my lord and father, madam,	~1\strandj.txt

Figure 29. Concordance extract for *I pray you*: Shakespeare's plays

N	Concordance		File
1		And when you have done so, I pray you remove your court further from m	1\ncalex.txt
2	will bind me much to you.	I pray you do not say so sir.	ncanhum.txt
3	u come away to dinner:	I pray you come hither.	ncmuced.txt
4	you lend me your dwarf.	I pray you, take him. Your hopes, sir, are li	1\ncvolp.txt
5	rch for.	Come I pray you, and be circumspect.	1\nholdc.txt
6	n their proper heads.	I pray you give them leave Madam, this sp	1\nhtami.txt
7	ajors under the Suns. Hark you me, Kings:	I pray you now, good Kings, leave your whi	\nhwelsh.txt
8	up of my M. Nag.	Why I pray you let us go before, Whilst he stays	r~1\ntaof.txt
9	n to the contrary. Sir, here is his passport,	I pray you sir, we have done him wrong.	1\ntspant.txt
10	re to blame else, And out of fashion much.	I pray you lead Sir. After	1\ntwbw.txt

Figure 30. Concordance extract for *I pray you*: other contemporaneous plays

In general, my data shows that *I pray you* is used in similar ways in both corpora. Characters use *I pray you* to mitigate requests in a variety of situations which are routine in the course of plays: for example, requests to speak, listen, come, go, give, take, or provide information. It is interesting that these relatively ordinary requests in Shakespeare's plays are paired with the sense of supplication of the complex form *I pray you*. Possibly this is to achieve a hyperbolic effect, which exaggerates the feelings of the speaker towards the addressee (adding emotion to the scene, and making the play more entertaining for the audience). I investigated whether or not this appears to be associated with particular kinds of contexts or situations, firstly by looking at the distribution of I PRAY YOU in both corpora. A dispersion plot for the *SDC* is shown in Figure 31 on the next page.

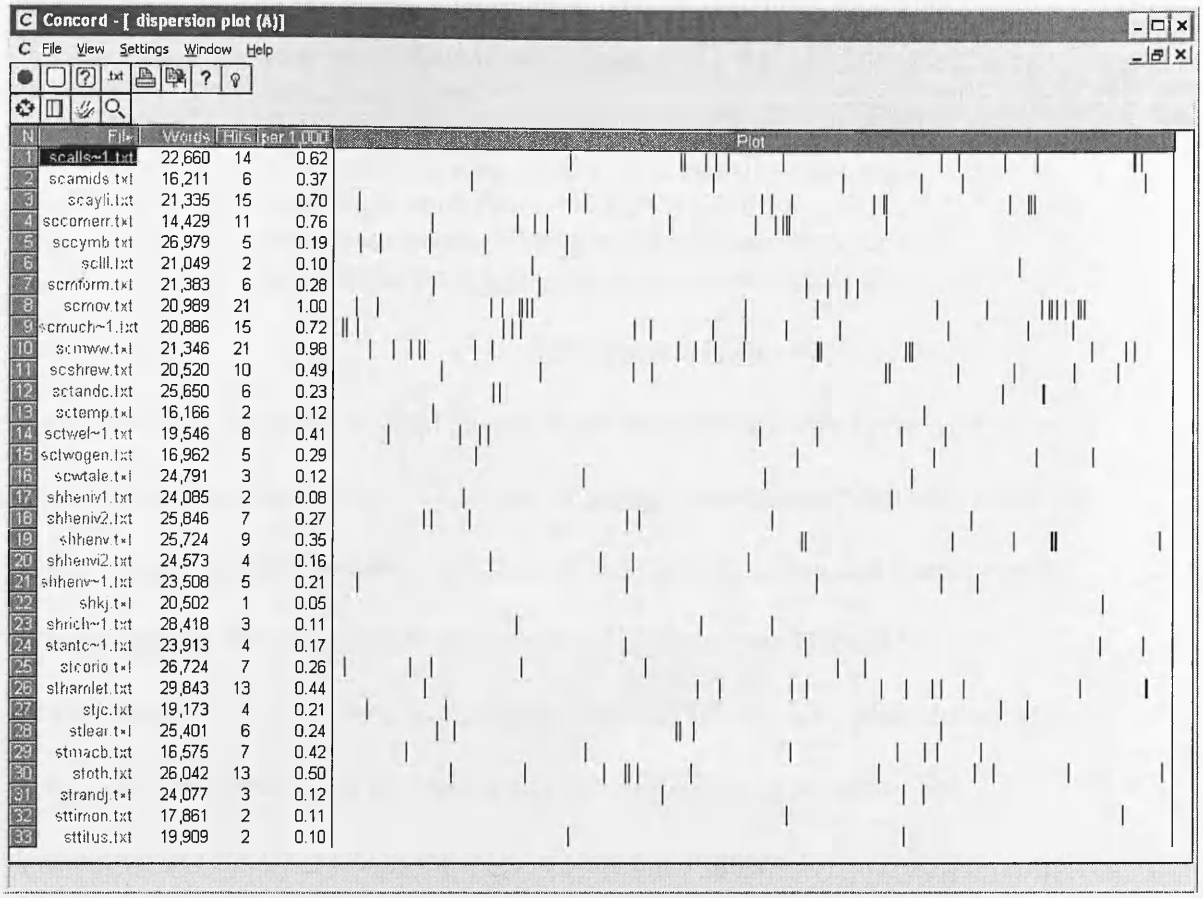


Figure 31. Dispersion plot for *I pray you* in the *SDC*

Figure 31 shows that there is a slightly greater density of I PRAY YOU in Shakespeare's comedy plays, in the top third of the plot, less in the histories below them, and still less in the tragedies at the bottom. This may be due to relatively more interaction in comedy, which would provide more potential for requesting, although it must also be remembered that the comedy sections of the corpora are larger in size than those of the other two genres (see 4.4), and they therefore contain more speech acts overall. There are few concentrations of I PRAY YOU in Figure 31. The largest, near the end of the history play *Henry V*, is due to its use by one particular male character, the Welsh army captain Fluellen, in a scene in which he beats the soldier Pistol and makes him eat a leek which he has mocked Fluellen for wearing. An extract is shown in example (29) on the next page.

(29) Fluellen: Eat, I pray you: will you have some more sauce to your leek? there is not enough leek to swear by.
Pistol: Quiet thy cudgel: thou dost see I eat.
Fluellen: Much good do you, scald knave, heartily. Nay, pray you, throw none away; the skin is good for your broken coxcomb. When you take occasions to see leeks hereafter, I pray you, mock at 'em; that is all.

Shakespeare, *Henry V*, V:i (*SDC*)

In example (29), the sense of humble supplication associated with *I pray you* is clearly not genuine, since Fluellen is in a position of greater power than Pistol and is forcing him to comply with the request to eat the leek, i.e. with an action that humiliates the addressee rather than the speaker. Fluellen's use of *I pray you* is therefore sarcastic, used to convey "mock politeness" (Culpeper's 1996:356 concept), which here creates humour: the encounter with the leek in this part of the play is comedic. The distribution of I PRAY YOU in the *NDC* is shown in Figure 32.

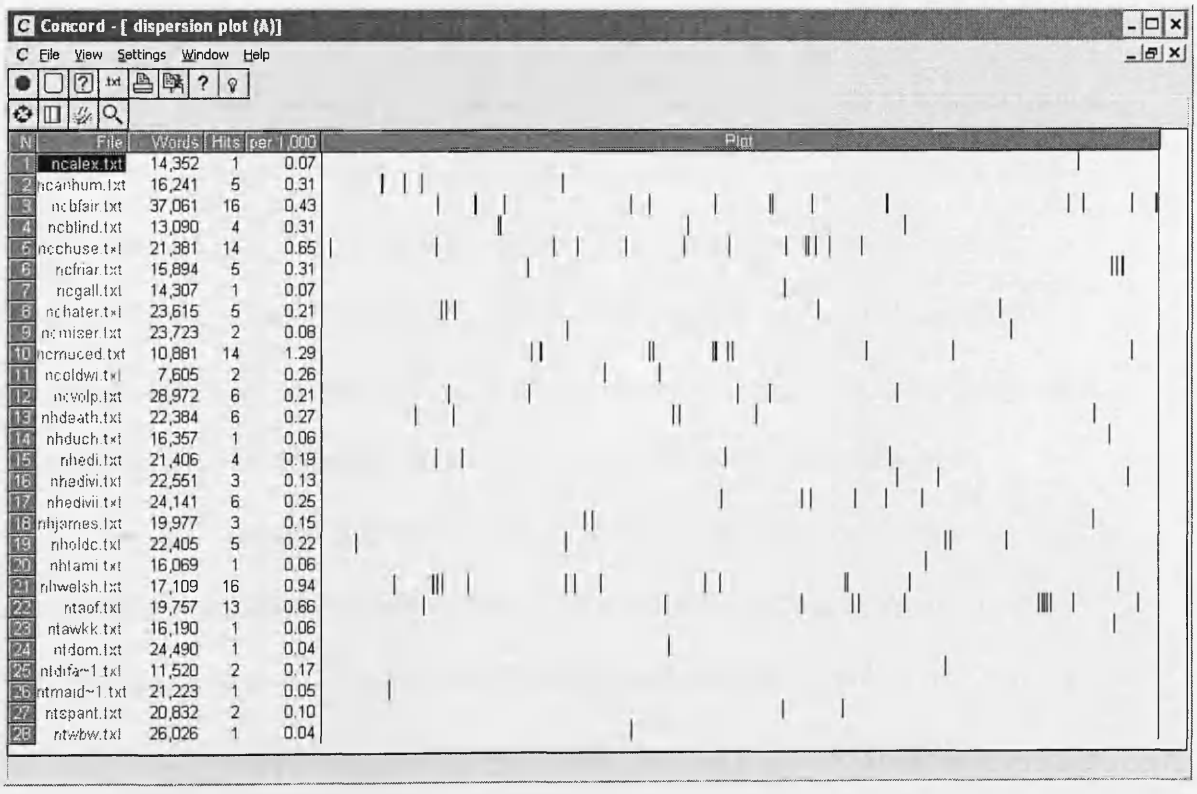


Figure 32. Dispersion plot for *I pray you* in the *NDC*

Though Figure 32 shows that the distribution of I PRAY YOU is clearly sparser in the other contemporaneous plays (as would be expected by its keyness in Shakespeare's plays), it is dispersed in a fairly similar pattern. The largest concentration, mid-way through the comedy *Mucedorus*, is accounted for by the use of *I pray you* by two male characters: Mouse, who is a clown, and his master Segasto. An extract is shown in example (30) below. Mouse wants Segasto to come and have dinner, but Segasto wants some information from Mouse concerning an errand on which he has previously been sent. Each tries to persuade the other to his will.

(30) Mouse: I pray you come away to dinner:
Segasto: I pray you come hither.
Mouse: Here's such a do with you, will you never come.
Segasto: I pray you sir what news of the message I sent you about

Anonymous, *Mucedorus* (*NDC*)

As in example (29) from Shakespeare's *Henry V*, in (30) the repetition of *I pray you* by Segasto creates a comedic effect of mock politeness when addressed to his servant. It would not be socially necessary for an employer to mitigate requests to a servant using the force of supplication conveyed by *I pray you*, so this clearly has a stylistic effect. Less complex forms of the pragmatic marker *pray* are found to be used between employers and servants in apparently conventional ways in Lutzky and Demmen (forthcoming), however, in which we also note the use of *I pray you* in an exaggerated, mock-polite way for humorous effect in Jonson's comedy *Bartholomew Fair*.

The above investigations from the corpora taken in their entireties are interesting, because the clusters of I PRAY YOU which stand out in the distribution plots in Figures 31 and 32 are accounted for by male dialogue in both sets of plays. I would not have expected this, since the findings from my previous research, mentioned in 2.5.4, show that I PRAY YOU is associated particularly with female dialogue in Shakespeare's plays, where it is the most key 3-word cluster when

compared to male dialogue (Demmen 2009:98-99). This keyness is located particularly in the comedy plays, which does tie in with the general pattern of distribution for both corpora shown in the plots above. Accordingly, I extracted key cluster results by gender and genre for both corpora, to see how these compare to the indications from the distribution plots.

The gender breakdowns confirm that I PRAY YOU is key in the female Shakespearean dialogue when compared to male Shakespearean dialogue, and also when compared to the other contemporaneous female dialogue as a reference corpus. When the gender sections of the corpora are analysed by genre, the keyness is again traced to the comedy plays in Shakespeare's plays. In the other contemporaneous plays, I PRAY YOU is not key (a) in female dialogue when compared to male dialogue, or (b) in male Shakespearean dialogue when compared to other contemporaneous male dialogue. By way of comparison with my present findings, in Lutzky and Demmen (forthcoming) we find *I pray you* to be only slightly more frequent in female dialogue (1.54 instances per 10,000 words for females to 1.40 for males). When I examined I PRAY YOU in my present data by genre only (without the gender variable), it is key in the Shakespearean comedies and tragedies (compared to the parallel genre sections of the other contemporaneous plays), but not in histories. An analysis of internal genre variation in each corpus shows that in Shakespeare's plays, I PRAY YOU is over-represented in comedy (compared to history and tragedy dialogue combined as a reference corpus), and under-represented in history plays. In the other contemporaneous plays, I PRAY YOU is under-represented in tragedy (compared to history and comedy combined), but not key otherwise.

The quantitative results above indicate that *I pray you* is a style feature which Shakespeare particularly favours in the dialogue of women in comedy plays (which

contrasts interestingly with his over-use of the more deferential *I beseech you* in male dialogue, discussed above). However, the concordance data for the gender and genre breakdowns of I PRAY YOU do not reveal anything particularly illuminating, other than to confirm that it is often used in making requests, which are common in both sets of plays, as noted above. In Demmen (2009:104-105) I suggest that Shakespeare's preference for including *I pray you* in female dialogue may be a representation of women using relatively more mitigation in requesting than men, linked to the lower social status of women compared to men in the Early Modern period. The absence of a similar finding in the other contemporaneous plays does not support this, however, because if it was socially conventional for women to behave in more humble ways in making requests, it would be strange for Shakespeare to be alone in representing them as doing so (in drama). I now consider some other possible explanatory factors.

The diachronic research into *pray* in EModE drama comedy carried out in Lutzky and Demmen (forthcoming) shows that the complex form *I pray you* declines in use much earlier than the less complex forms, and is not found in our data after 1639. T. Walker (2007:271), too, notes that complex *pray* forms decline after 1639 in both EModE drama comedy and courtroom trial data. Shakespeare's relatively frequent use of I PRAY YOU compared to his peers, in my present data, may therefore be due to a preference for a language form which is becoming rather old-fashioned and falling out of favour with younger speakers. This would fit with the findings of Craig (2012) and Ingram and Ingram (2012), mentioned in 2.4.3, who argue that Shakespeare's language style tends to be more conservative than that of his contemporaries, particularly the younger ones. It would not explain why it is relatively highly frequent in the comedy genre, though, bringing me back to the above point

(made just after Fig. 31) about the relatively greater amount of interaction in comedies, which is a more likely explanation.

I do not explore other *pray* forms in this study, but the fact that they do not occur as key in any of my data leads me to suggest that *I pray you* is a special case among the group of *pray* pragmatic markers, where its connotations with extreme humility and supplication (argued by Culpeper and Archer 2008) are harnessed for other stylistic effects (exaggerated deference and dramatic hyperbole). The clusters of I PRAY YOU which were pinpointed in Figures 31 and 32 above were few, but they usefully led to findings of the use of *I pray you* as a device for mock politeness which Shakespeare and other playwrights made use of. This may be a particular feature of the register of EModE drama which Shakespeare opts to exploit to a greater extent than other dramatists, especially in the dialogue of women. He also creates a hyperbolic effect in male dialogue, through the over-use of the more formal and deferential cluster I BESEECH YOU. There is some evidence that the stylistic potential of I PRAY YOU is not limited to drama, however. In data from two EModE trials, Włodarczyk finds that in one *I pray you* "is straightforwardly deferential", while in the other there are cases in which it is "used ironically or in a challenging way" with "layered pragmatic and rhetorical functions" (2007:122).

Finally, I cannot rule out the possibility that the keyness of I PRAY YOU is associated with a difference in the number of speech acts of requesting in the two corpora, particularly in the dialogue of women and in comedy plays, since that is where the keyness is traced to in Shakespeare's plays. This factor is also noted by Lutzky and Demmen (forthcoming), but counting the number of requests in both corpora, as did Culpeper and Archer (2008:76), was beyond the resources of this study. However, the greater presence of other deferential clusters in Shakespeare's

plays noted in 6.3, plus the evidence of more use of *beseech* (as a single word and in clusters, in this and the previous section), lend weight to the argument that *I pray you* is a Shakespearean style feature and not merely a by-product of requesting. The reason why Shakespeare favours *I pray you* in the dialogue of female characters is intriguing, though not entirely clear.

Apart from the *beseech* clusters and I PRAY YOU, several other key clusters in Table 24 (at the start of this section) also pertain to politeness and deference, and potentially tie in with the evidence that Shakespearean dialogue contains relatively more deferential language than that in the other contemporaneous plays. These are the Interpersonal: Speech-act-Related: Vocative MY GOOD LORD, the Directive politeness clusters FARE YOU WELL and FARE THEE WELL and the Modalising cluster SO PLEASE YOU. However, FARE YOU WELL and FARE THEE WELL also have a role in precipitating the movement of characters on and off stage (as argued with regard to the lockword FAREWELL in 7.2). Therefore, these may simply be stagecraft formulae which Shakespeare made more use of than his peers.

Finally amongst the key clusters in Table 24, the single result WHAT IS THE in the Textual: Discoursal: Question category is notable. The concordance data shows that half the 58 instances are part of the longer formula *what is the matter*, and that it occurs in all three genres of Shakespeare's plays, but with a notable concentration in the tragedy *Othello*. The function of this question is to elicit information between characters on stage, and it is probably a strategy which Shakespeare favoured to help both the audience and the actors in engaging with the play, as I will illustrate with some examples. Culpeper and McIntyre (2006:773) claim that playwrights embedded strategies in EModE dramatic dialogue to help counteract the inattention of audience members, who would have been talking, eating, or otherwise distracted while the

performance went on. In example (31), the question *what is the matter* from Gratiano sets up an opportunity for the addressee, Emilia, to summarise the plot situation (as she sees it), in terms of who has said and done what to whom, and what needs to happen next. This would have filled in any audience members who had failed to grasp it earlier on.

(31) Gratiano: What is the matter?
Emilia: Disprove this villain if thou be'st a man:
He says thou told'st him that his wife was false.
I know thou didst not, thou'rt not such a villain,
Speak, for my heart is full.

Shakespeare, *Othello*, V:ii (*SDC*)

The same question is also regularly used to set up opportunities for characters to point out things happening on stage which are important for the audience not to miss (through inattention, or perhaps lack of close proximity to the stage). In the series of examples below, again from *Othello*, Montano responds to *what is the matter* by highlighting a mortal injury (32), Bianca uses the question to allude to the sound of a cry (33), and finally Othello responds to the question by drawing attention to the fact that he is holding a weapon (34).

(32) Othello: What is the matter here?
Montano: 'Zounds! I bleed still; I am hurt to the death.

Shakespeare, *Othello*, II:iii (*SDC*)

(33) Bianca: What is the matter, ho? who is 't that cried?

Shakespeare, *Othello*, V:i (*SDC*)

(34) Gratiano: What is the matter?
Othello: Behold! I have a weapon;

Shakespeare, *Othello*, V:ii (*SDC*)

Not only do the questions in examples (32) to (34) serve to highlight what the audience needs to have noticed, they also provide what Aston and Savona (1991) term

"intra-dialogic" stage directions for the actors. These, they argue, are stage directions which are part of the dialogue itself, since "extra-dialogic" stage directions (those annotated to the script) were relatively few in EModE plays (Aston and Savona (1991:75-78; see also Culpeper and McIntyre 2006:776-777). The intra-dialogic stage directions would have cued the actors to the physical behaviours they needed to manifest at that point in the play. This would have been essential, since at the time the plays were written, dramatists gave the actors only the portions of the script which contained the dialogue of their own character role(s), with a few words of the previous speaking turns by other actors as cues (see the summary in Culpeper and Demmen 2011, and more detailed discussions in Stern 2000, 2004). The question fragment *what is the* therefore seems to be a useful dialogic strategy which Shakespeare favoured relatively more than his contemporaries. Skills in deploying such strategies to counteract the limitations of a potentially distracted audience, and to aid actors who did not know the whole play in advance of performance, would of course have contributed to the quality of the drama, and to the success of the plays (and the playwright).

In the next section, I analyse the key semantic domains in the *SDC*, to see what kinds of concepts are used relatively more or less in Shakespeare's plays compared to plays by other contemporaneous dramatists.

8.4 Key semantic domains in Shakespeare's plays

The top 10 positive key semantic domains and negative key semantic domains in Shakespeare's plays are respectively shown in Tables 25 and 26. Raw frequencies and log-likelihood ("LL") values are shown for both corpora, and the results are displayed in descending order of keyness (LL value). The three most frequently-occurring

examples in the *SDC* are shown below each category label, in italics. Note that the USAS tool categorises certain compounds as single items (e.g. *for ever* and *all night* in category T1.3+, ranked 10th in Table 26).

Table 25. Top 10 rank-ordered positive key semantic domains in Shakespeare's plays (minimum frequency=200; LL=27)

Rank	**Positive key domains**	***SDC***	***NDC***	**LL**
1	Personal names (Z1) e.g. *will, York, Warwick*	13,470	9,253	595.15
2	Speech: communicative (Q2.1) e.g. *speak, said, told*	3,865	2,813	119.12
3	Detailed (A4.2+) e.g. *very, certain, particular*	657	322	101.24
4	Grammatical bin (Z5) e.g. *the, and, to*	203,384	188,307	78.31
5	Anatomy and physiology (B1) e.g. *heart, hand, blood*	11,887	10,045	78.14
6	Speech acts (Q2.2) e.g. *say, tell, name*	11,153	9,526	61.13
7	Calm (E3+) e.g. *peace, gentle, patience*	1,667	1,187	59.42
8	Green issues (W5) e.g. *nature, natures, polluted*	395	209	49.51
9	Degree: maximizers (A13.2) e.g. *most, all, altogether*	1,105	763	47.41
10	Living creatures: animals, birds, etc. (L2) e.g. *horse, dog, lion*	3,440	2,756	46.01

Table 26. Top 10 rank-ordered negative key semantic domains in Shakespeare's plays (minimum frequency=200; LL=27)

Rank	Negative key domains	*SDC*	*NDC*	LL
1	Unmatched (Z99) e.g. *csar, didst, canst*	12,906	16,194	549.81
2	Other proper names (Z3) e.g. *which, more, who*	1,412	2,680	463.68
3	Sensory: sight (X3.4) e.g. *see, seen, saw*	2,682	3,471	143.55
4	Linear order (N4) *then, first, last*	3,563	4,298	109.31
5	Money: affluence (I1.1+) e.g. *rich, wealth, prodigal*	307	549	81.66
6	Time: future (T.1.1.3) e.g. *will, shall, 'll*	11,154	11,871	70.94
7	Exclusivizers/ particularizers (A14) e.g. *only, just, alone*	512	749	57.11
8	Stationary (M8) e.g. *stay, sit, still*	925	1,209	53.01
9	Entire; maximum (N5.1+) e.g. *all, any, every*	5,334	5,813	50.55
10	Time period: long (T1.3+) e.g. *long, for ever, all night*	496	707	48.2

Some of the semantic domains in Tables 25 and 26 point to potentially interesting contrasts in the concepts used most statistically frequently in the two corpora, a number of which relate to metaphors, as was the case with the locked domains in 7.4. However, others are problematic, either because of inappropriate categorisation by the USAS tool, or because they can be explained by the nature of the texts themselves, and I discuss these briefly first.

The presence of "Personal names" as the most over-used semantic domain in the *SDC* (Table 25) and "Unmatched" as the most under-used (Table 26), arise from the lower levels of standardisation of spelling in the early extant *NDC* texts compared to the modernised *SDC* texts (discussed in 5.4 and mentioned in 8.1 and 8.2). USAS successfully identifies relatively fewer personal names in the *NDC*, because most name spellings have not been regularised, and relatively more "Unmatched" word

forms overall, because more variants remain after *VARD 2* was applied to regularise the spelling. It may nevertheless be the case that Shakespeare uses more proper names than his contemporaries; at the end of 7.2 I noted that this could be an explanatory factor in the positive keyness of the pronouns HE and HIM and the negative keyness of the vocative MISTRESS in the *SDC*. Unfortunately, this would be difficult to investigate further because of the huge number of names in each corpus.

The relatively under-used category of "Other proper names" in the *SDC* (ranked second in Table 26) is also unreliable. This is because some words which are common nouns in the *NDC* have initial capitals, (e.g. *More* and *Sun*), which apparently cause them to be tagged as proper nouns by the USAS tool, falsely inflating the size of the domain in the *NDC*. The words in the "Grammatical bin" are function words, and their ranking of 4th in the positive key domains in Table 25 indicates that they are over-represented in Shakespeare's plays. However, there was strong evidence of similar types and frequencies of function words in both corpora in the normalised frequencies examined in 6.2. The reason for the keyness of this semantic domain category therefore appears to be the lower level of standardisation in the *NDC* texts, particularly of apostrophes and contracted forms.

I now look at selected results from the more reliable key semantic domains, beginning with concepts which are used relatively more often in Shakespeare's plays than in the other contemporaneous plays (in Table 25). The "Calm" domain, ranked 7th, is reliable in that it contains words and concepts which are associated with calmness. However, is not very cohesive, and the grouping of concepts into this domain does not really aid stylistic analysis, because the functions of the most frequently-occurring items are very different and outweigh the semantic similarity. I will illustrate this briefly.

Although some instances of the most frequently-occurring words in the "Calm" domain (*peace*, *gentle* and *patience*) are sometimes used to refer directly to calmness, in my data they are more often used to perform particular pragmatic functions. The word *peace* is most frequently used as a request for an addressee to be quiet and/or settle down. Some examples are given in the concordance extract from the *SDC* in Figure 33.

1.	I shall never begin if I hold my	peace	. Good , i' faith . Come , begin . W
2.	December , - For the love o' God ,	peace	! My masters , are you mad ? or what
3.	Here 's an over-weening rogue ! O,	peace	! Contemplation makes a rare turkey-
4.	light , I could so beat the rogue !	Peace	! I say . To be Count Malvolio ! Ah
5.	, rogue ! Pistol him , pistol him .	Peace	! peace ! There is example for it :
6.	! Pistol him , pistol him . Peace !	peace	! There is example for it : the lady
7.	rdrobe . Fie on him , Jezebel ! O ,	peace	! now he 's deeply in ; look how ima
8.	eeping , - Fire and brimstone ! O ,	peace	! peace ! And then to have the humou
9.	- Fire and brimstone ! O , peace !	peace	! And then to have the humour of sta
10.	n Toby , - Bolts and shackles ! O ,	peace	, peace , peace ! now , now . Seven

Figure 33. Concordance extract for *peace*: Shakespeare's plays

Peace, therefore, although being associated with calmness in Shakespeare's plays (Crystal and Crystal 2002:322), does not actually make reference to an existing state; it is a response to the opposite kind of state, in a frequently-occurring request.

The function of *gentle* is very different to that of *peace*. My data indicates that it is most often used in address as a politeness marker implying nobility, e.g. *gentle friend*, *gentle lady*, *gentle duke* (see also Crystal and Crystal 2002:197). Occasionally it is used on its own to imply intimacy (as a single term of address, *gentle* means "dear one" according to Crystal and Crystal 2002:8). There is a previously-identified Shakespearean style preference for the word *gentle*, in the computational stylistic research of Craig and Kinney (2009b:38), who find it is relatively over-used by Shakespeare in comparison to other contemporaneous playwrights. Although *gentle* does not appear in the top 20 keywords (in 8.2), its major numeric contribution to the keyness of the "Calm" domain tends to support Craig and Kinney's findings.

Patience, the third most frequent word in the Calm key domain in Shakespeare's plays, has positive connotations. Some examples are given in the concordance extract from the *SDC* in Figure 34, which shows it being described as a virtue (line 10), associated with love (line 4) and kissing (line 8), and described as a "goddess" (line 9).

1.	ress : be moved , be moved . Have	patience	, gentle Julia . I must , where is
2.	les on equal mates , And think my	patience	, more than thy desert , Is privil
3.	If you be she , I do entreat your	patience	To hear me speak the message
4.	s I see and hear ! Love , lend me	patience	to forbear awhile . O ,
5.	on blush , and tyranny Tremble at	patience	. You , my lord , best know , -
6.	ord , Who is lost too : take your	patience	to you , And I 'll say nothing .
7.	as my tale Now seems to it . Your	patience	this allowing , I turn my glass
8.	that hand of yours to kiss . O ,	patience!	The statue is but newly fixed ,
9.	ou may chance to burn your lips .	Patience	herself , what goddess e'er she
10.	o see the battle . Hector , whose	patience	Is as a virtue fixed , today was

Figure 34. Concordance extract for *patience*: Shakespeare's plays

While there is an association between patience and calm in Shakespearean drama (according to Crystal and Crystal 2002:321), the fact that characters are asking one another to be patient actually implies that calmness is not the present state, hence the reason for mentioning it. Tissari (2010b:310) points out that "'calmness' is necessary for social interaction to be successful (that is why people sometimes need to *calm down*)".

The three most frequent words in the "Calm" domain are the same in the *NDC* as in the *SDC* (*peace*, *gentle* and *patience*), and the concordance data indicates that they are used in similar ways. Extracts for *peace* are shown in Figure 35 and for *patience* in Figure 36 (on the next page).

1.	xander . You can neither brook this	peace	, nor my pleasure , be of good cheer
2.	you a receipt for this presently .	Peace	Lemot , they say the young lord Dows
3.	l save your honesty for this once .	Peace	, a plague on you , peace ; but wher
4.	is once . Peace , a plague on you ,	peace	; but wherefore asked you how I did
5.	; It is true , she is A Justice of	Peace	his wife , and a Gentlewoman of the
6.	d 't be but for conservation of the	peace	. Mary gip , goody she-Justice , Mis
7.	to it , for the t' other remnant .	Peace	, Urs , peace , Urs , they 'll kill
8.	he t' other remnant . Peace , Urs ,	peace	, Urs , they 'll kill the poor Whale
9.	will conclude briefly--- Hold your	peace	, you roaring Rascal , I 'll run my
10.	keep it during the Fair , Bobchin .	Peace	, Numps , friend , do not meddle wit

Figure 35. Concordance extract for *peace*: other contemporaneous plays

1.	ast . Go to then , rest here with	patience	, and be confident in my trust , o
2.	could be made a Cuckold with more	patience	, than endure this . We . For we s
3.	han a knave of three and twenty ,	Patience	be my Buckler , As not to file my
4.	nce where 's bounties throng Give	patience	to my soul , inflame my tongue . G
5.	I am very poor and very patient ,	Patience	is a virtue : would I were not vir
6.	ter no . O my dear more have some	patience	, Aye sir , have patience , and se
7.	ve some patience , Aye sir , have	patience	, and see your father To rifle up
8.	lot ? This same will make me have	patience	, will it not ? This same is women
9.	bed it , commending the virtue of	patience	or forbearance , but yet you know
10.	erance to forbear drink so have I	patience	to endure drink , I 'll do as comp

Figure 36. Concordance extract for *patience*: other contemporaneous plays

The examples of *peace* from the *NDC* in Figure 35 seem less forceful than those in the *SDC* (Fig. 32), due to the absence of serial repetitions and exclamation marks which characterise the Shakespearean examples. However, as argued in 5.4, punctuation may be a result of the compositors' intervention, not a reflection of authorial style. *Patience* is often requested by one character from another in other contemporary plays (lines 1, 6 and 7, Fig. 36), and described as a virtue (lines 5 and 9).

The above analyses of the three most frequent words in the "Calm" category show that semantic similarity does not necessarily indicate similarity of function in dialogue (or, therefore, similar stylistic effects). Furthermore, the fact that *peace*, *gentle* and *patience* did not arise as individual keywords, in 8.2, suggests that there may not be a quantitatively significant difference in their use by Shakespeare and the other contemporary dramatists, although they are the most numerous single words in the "Calm" domain – apart from *gentle*, for which there is other evidence supporting

its over-use by Shakespeare (from Craig and Kinney 2009b). This highlights a limitation of the semantic domain method. Although it is useful in grouping words which might not occur as statistically significant on their own (argued by Rayson 2008:543 and borne out by the locked domain results concerning items of everyday social life in 7.4), there is also the potential for it to include high-frequency words which are not on their own statistically significant, but whose presence boosts the statistical significance of a particular semantic group and thereby amplifies its apparent importance in the texts. In my data *peace* and *gentle* could be more fruitfully analysed either as individual words, or grouped with other words of similar function, instead of similar semantic meaning. Crystal and Crystal (2002:8) classify *gentle* among a group of other address forms in Shakespeare's plays, and they also identify a small but distinctive group of words and phrases which are "attention signals", including *peace!*, which they claim is a particularly important function in dramatic interaction (2003:26). Address forms and attention signals would be worth exploring further in the other contemporaneous plays, in future research.

Other positive key domains in Table 25 suggest that there is relatively more talk about communication itself among Shakespeare's characters than those in the other contemporaneous plays. The concepts in the "Speech: communicative" and "Speech acts" domains overlap to some extent, both containing a mix of reported speech and words relating to communication. Some examples are given in the concordance extract from the "Speech: communicative" domain in Figure 37 (on the next page).

1.	thy pains ; for I can sing And	speak	to him in many sorts of music T
2.	ill undo you : I heard my lady	talk	of it yesterday ; and of a fool
3.	e unprofited return . Say I do	speak	with her , my lord , what then
4.	of my love ; Surprise her with	discourse	of my dear faith : It shall bec
5.	r : I can tell thee where that	saying	was born , of , 'I fear no colo
6.	e thee with leasing , for thou	speakest	well of fools ! Madam , there i
7.	oung gentleman much desires to	speak	with you . From the Count Orsin
8.	etch him off , I pray you : he	speaks	nothing but madman . Fie on him
9.	nd young fellow swears he will	speak	with you . I told him you were
10.	ars he will speak with you . I	told	him you were sick : he takes on

Figure 37. Concordance extract for Speech: communicative (semantic domain Q2.1): Shakespeare's plays

The "Speech acts" domain is broad, and some of the contents are problematic.

Examples are shown in Figure 38, and discussed below.

1.	Sir Andrew ! Bless you , fair	shrew	. And you too , sir . Accost ,
2.	esire better acquaintance . My	name	is Mary , sir . Good Mistress M
3.	' is , front her , board her ,	woo	her , assail her . By my troth
4.	nt her , board her , woo her ,	assail	her . By my troth , I would not
5.	that does harm to my wit . No	question	. An I thought that , I 'd fors
6.	ion . An I thought that , I 'd	forswear it	. I 'll ride home tomorrow ,
7.	hat have mended my hair ? Past	question	; for thou seest it will not cu
8.	The count himself here hard by	woos	her . She 'll none o' the count
9.	s , nor wit ; I have heard her	swear	it . Tut , there 's life in it
10.	ur or my negligence , that you	call	in question the continuance of

Figure 38. Concordance extract for Speech acts (semantic domain Q2.2): Shakespeare's plays

In the above cases, *shrew* (line 1, Fig. 38) is categorised by USAS as a speech act, probably because it can be a short form of the verb *beshrew*, which means to "curse" (Crystal and Crystal 2002:40). However, here it is a noun describing a woman, used as a term of address, not a speech act. The word *question* (lines 5 and 7, Fig. 38) is used idiomatically to refer to a hypothetical speech act of questioning, but its function is to boost or emphasise the speaker's point.

While Shakespeare's characters use relatively more concepts surrounding qualities of calmness and aspects of speech and communication, the negative semantic domain "Sensory: sight" in Table 26 indicates that they talk relatively less about what they have seen. The most frequent word in it, *see*, was noted as being less frequent in

Shakespearean drama in 6.2, and occurring as a negative keyword in 8.2. Furthermore, concepts to do with time (including the future, periods of time and lengths of time) and money are used relatively less frequently in Shakespeare's plays. The concordance data shows that these concepts are used in similar ways in dialogue by Shakespeare and his peers, just apparently to a lesser extent by Shakespeare. The "Money: affluence" domain is mainly populated with the words *rich* and *wealth* in both corpora, and it is interesting that this, together with the other everyday concept of time, do not occur to a similar extent in both sets of plays. I would have expected them to arise amongst the domains which lock across both corpora, together with other everyday concepts such as weather and furniture, discussed in 7.4 above. This is partly from intuition, but also because Baker (2011) finds *money* to be a stable high-frequency word in British English (albeit over more recent times), as stated in 2.6.2.

The data in some of the other positive key domains in Table 25 indicates that Shakespeare makes relatively greater use of other concepts, however, particularly those involving the human body, nature or animal life. These concepts readily lend themselves to metaphors, some of which were also noted in the data from the most frequent domains in each corpus (in 6.4) and in the locked domains (in 7.4), as I now illustrate briefly. Examples of body concepts used in Shakespeare's plays are shown in the concordance extract in Figure 39.

1.	, that did rescue me . That	face	of his I do remember well ; Y
2.	eet , That very envy and the	tongue	of loss Cried fame and honour
3.	young nephew Titus lost his	leg	. Here in the streets , despe
4.	cies Whom thou , in terms so	bloody	and so dear , Hast made thine
5.	ude sea 's enraged and foamy	mouth	Did I redeem ; a wrack past h
6.	s as fat and fulsome to mine	ear	As howling after music . Stil
7.	faithfull'st offerings hath	breathed	out That e'er devotion tendered !
8.	Why should I not , had I the	heart	to do it , Like to the Egypti
9.	ill I tear out of that cruel	eye	, Where he sits crowned in hi
10.	o love , To spite a raven 's	heart	within a dove . And I , most

Figure 39. Concordance extract for Anatomy and physiology (semantic domain B1): Shakespeare's plays

Some of the body concepts in Figure 39 are used in metaphorical ways, e.g. "the tongue of loss Cried" (line 2) and "sea's enraged and foamy mouth" (line 5). Shakespeare's innovative use of body concepts to powerful dramatic effect has been noted in other studies. For example, regarding *Titus Andronicus*, Tricomi argues that:

> In a play preeminently concerned with the mutilation of the human body, *Titus* makes nearly sixty references, figurative as well as literal, to the word 'hands' and eighteen more to the word 'head', or to one of its derivative forms. [...]
>
> By shackling the metaphoric imagination to the literal reality of the play's events, the tragedy strives for an unrelieved concentration of horrific effect. (2004 [1974]:226, 237)

The positive key domain "Green issues" is almost exclusively populated by one word: *nature*. The concept of nature is often personified by Shakespeare, some illustrations of which are given in the concordance extract in Figure 40.

1.	e in loving be ; And the blots of	Nature's	hand Shall not in their
2.	tretched so far , would have made	nature	immortal , and death
3.	ot politic in the commonwealth of	nature	to preserve virginity . Loss
4.	in it ; 't is against the rule of	nature	. To speak on the part of
5.	as a desperate offendress against	nature	. Virginity breeds mites , ?
6.	The mightiest space in fortune	nature	brings To join like likes ,
7.	ear'st thy father 's face ; Frank	nature	, rather curious than in haste , H
8.	e out With several applications :	nature	and sickness Debate it at their le
9.	when I was young : If ever we are	nature's	, these are ours ; this
10.	born : It is the show and seal of	nature's	truth , Where love 's strong pa

Figure 40. Concordance extract for Green issues (semantic domain W5): Shakespeare's plays

Many of the 3,440 words which populate the positive key domain "Living creatures: animals, birds, etc." in Shakespeare's plays are also used to describe one thing in terms of another, either in metaphors or similes. Some examples are shown in Figure 41 on the next page. Metaphors include "She's a beagle, true-bred" (line 1), used by a male character to describe a female character, and "here comes the trout that must be caught" (line 5), used by a female character to describe a male character. Amongst the similes are "like a worm i' the bud" (line 3) and "as rank as a fox" (line 8).

1.	she 's a good wench . She 's a	beagle	, true-bred , and one that ador
2.	e in the constant image of the	creature	That is beloved . How dost thou
3.	, But let concealment , like a	worm	i' the bud , Feed on her damask
4.	. To anger him we 'll have the	bear	again ; and we will fool him bl
5.	hou there : for here comes the	trout	that must be caught with tickli
6.	how he jets under his advanced	plumes!	'Slight , I could so beat the
7.	ment have we here ? Now is the	woodcock	near the gin . O , peace !
8.	is , though it be as rank as a	fox	. M , Malvolio ; M , why , that
9.	fools are as like husbands as	pilchards	are to herrings-the husband
10.	e the haggard , check at every	feather	That comes before his eye . Thi

Figure 41. Concordance extract for Living creatures: animals, birds, etc. (semantic domain L2): Shakespeare's plays

Again, it is important to stress that dialogue authored by the other contemporaneous dramatists also features metaphors in which the source domain is the body, nature or animal life. However, the fact that these emerge as key semantic domains, coupled with the evidence for constituent lexical items being used in the context of metaphors (from the concordance data), suggests that Shakespeare uses them to a greater extent than his contemporaries. I argued in 7.4 that Shakespeare and other dramatists make use of the device of pathetic fallacy, using personification to create special effects in dramatic dialogue. However, the extent to which Shakespeare does so may be greater, given that Hope states that "[a]gency and process are the properties Shakespeare most characteristically bestows on the things he writes about", which, he argues, contributes to "the fluid, rather than static nature of Shakespeare's use of language" (2004:13-14). More detailed research into the most frequently-used concepts in the key domains (such as *nature*) would be useful, and could be linked to other studies of Shakespeare's language. For example, B. Busse (2006:332) finds that "natural phenomena" are used as vocatives in Shakespearean drama, as noted in 7.4, and it would be worth exploring whether Shakespeare makes greater use than his peers of vocatives comprising nature and animal metaphors.

I will now briefly summarise the remaining key domains in Tables 25 and 26. The positive key domain "Degree: Maximisers" (ranked 9th in Table 25) contains

mainly instances of the word *most*, which Shakespearean characters use to boost what they are saying (e.g. *most strange*, *most dear lady*). There are also some instances of *all*, which functions in a hyperbolic way, again to boost or add emphasis. Example (35) shows *all* used in the context of *the world*, which was noted as a lockword with a hyperbolic function in 7.2.

(35) Gentleman 3: if all the world could have seen't, the woe had been universal.

Shakespeare, *The Winter's Tale*, V:ii (*SDC*)

All is also sometimes categorised by USAS into the "Entire: maximum" domain (N5.1+), which is negatively key in Shakespeare's plays, but the concordance data shows that some instances are nevertheless used in metaphorical contexts, to add emphasis to what the speaker is saying. Some re-categorisation would be required in order to get a more accurate picture, but it does appear from the results that Shakespeare's characters boost their claims using *most* or *all* to a greater extent than characters in other contemporaneous plays. Conversely, Shakespeare's characters make relatively less use of words which order events and information, such as *then*, *next* and *before* (as revealed by the negative key domain "Linear order", ranked 4th in Table 26) and relatively less use of words such as *only* and *alone*, in the "Exclusivizers/particularizers" domain (ranked 7th in Table 26). The negative key domain "Stationary" (ranked 8th in Table 26) shows that Shakespeare's characters talk less about staying and sitting than those in other contemporaneous plays (*stay* and *sit* were the most frequently occurring words in this domain, in both corpora). This would also be worth exploring further in future research.

Although not without some problematic categories, the key semantic domain results provide an overall picture of the concepts which Shakespeare's characters tend to mention relatively more or less often, compared to those in plays by other

dramatists of the period, as well as some interesting leads which might usefully be followed up in future studies. It is important to emphasise here that it is a question of degree, not of absolute contrasts between concepts used in the two sets of plays. For example, Shakespeare and the other dramatists use body and nature metaphors, but the data suggests that Shakespeare makes relatively more use of them. The relatively lower incidence of concepts surrounding time and money in Shakespeare's plays is also potentially interesting. This could be related to the topics of the plays to some degree. However, the *NDC* was carefully balanced so that plays which were likely to be oriented around wealth, such as city comedies, were not over-represented, and pastoral themes such as those in some of Shakespeare's comedies were also included (as discussed in 4.3.2.2).

This concludes my analyses of the key results in Shakespeare's plays, when they are compared to the other contemporaneous plays. I now draw together the findings from this chapter, and provide an overall summary of the semantic domain analyses in the study.

8.5 Discussion and conclusions

8.5.1 Implications for Shakespeare's style and for the register of Early Modern English drama

Among the key results, some deferential language features have emerged which Shakespeare uses relatively more than the other contemporaneous dramatists. These have enlarged the initial picture suggested by the most frequent 3-word clusters in 6.3 (in which more of those in Shakespeare's plays had deferential associations). Shakespeare has been shown to have a stylistic preference for using particularly humble forms of requesting relatively more often than the other contemporaneous dramatists, specifically:

- the verb *beseech*, particularly in male dialogue (in 8.2 and 8.3); and
- the pragmatic marker *I pray you*, particularly in female dialogue (in 8.3).

With regard to beseeching, the evidence indicates that whereas the other contemporaneous dramatists tend to reserve it for relatively grave situations (such as for the mitigation of punishment), Shakespeare injects it into fairly ordinary situations such as coming or going. The stylistic effect is to overlay relatively ordinary requesting situations with the feelings and emotions that pertain to requests of much greater consequence, making them seem more dramatic. In this way, the language of Shakespeare's plays is more highly emotionally charged, and therefore arguably more exciting, than that of the other contemporaneous plays, although many of them share common themes and topics, and in other ways the language is similar (as demonstrated through the locked results in chapter 7).

With regard to *I pray you*, Shakespeare and the other contemporaneous playwrights use it in similar ways to exaggerate the humility in routine or ordinary requests, and also to create comedy through mock politeness (Culpeper 1996:356) through the sarcastic twisting of its sense of supplication (Culpeper and Archer 2008:76). It is, therefore, arguably a register feature in EModE drama, since it is not used exclusively by Shakespeare, although evidence from Włodarczyk (2007:122) indicates that its potential for stylistic effects extends beyond the drama register. My findings indicate that Shakespeare exploits *I pray you* much more often in comparison to his contemporaries in female characters' speech and in the comedy genre.

8.5.2 Corpus compatibility issues highlighted by the keyness analyses

Some possible problems in using modernised texts as the basis for the *SDC* were anticipated (in 3.2.2), but considered acceptable given that using an existing

Shakespeare corpus facilitated the construction of the new *NDC* for the study. From the keywords in 8.2 and the key clusters in 8.3, it is clear that results which contain punctuation, particularly grammatical contractions and possessives with apostrophes, and word forms that could occur as compounds in the Early Modern period, need to be scrutinised closely to see whether they reflect inconsistent orthography between the corpora. Some of the key semantic domains, in 8.4, were not reliable because of the lack of standardisation of personal names and other proper nouns in the *NDC*. These problems which affect the retrieval of reliable results cannot be overcome fully when using corpus linguistic methods with historical texts.

Spelling regularisation processes (discussed in 5.4) improve reliability, but complete standardisation could not be achieved with texts of this period without making what amount to editorial decisions (e.g. modernising any compound pronouns). Some compromises must therefore be accepted as a limitation in this kind of study. It is a pity that the spelling variation issues surrounding pragmatic markers AY, O and HA (8.2) mean that more detailed analyses would probably not be reliable, since this would be an interesting area of comparison between Shakespearean and other contemporaneous plays. However, as my discussions in 8.2-8.4 show, there were plenty of reliable key results from which fruitful stylistic analyses could proceed.

8.5.3 Evaluating the keyness method

The keyness results showed less evidence pointing to register features in EModE drama than the locked results in the previous chapter, as would be expected from their orientation to difference rather than similarity. Having noted a great many frequent clusters with the Ideational: Topical function of expressing states in both corpora (in 6.3), it was interesting to see that they are scarce when the clusters in both corpora are

compared statistically in the keyness analysis (in 8.3). This indicates that they effectively cancel one another out, through being fairly similar in type and frequency in both corpora. They can therefore be considered to be more akin to register features of EModE drama, since they represent a common set of referential building blocks which are used by Shakespeare and other dramatists to convey what is going on in the plays. The evidence from these two separate analyses of 3-word clusters in the corpora was combined in order to see this. Had the key clusters been looked at in isolation, the States clusters would simply not have been observed, and had the highly frequent clusters been examined on their own, their numerousness would have overshadowed other cluster results of greater stylistic interest.

Furthermore, it is worth noting that the verb *beseech*, on its own, does not occur as a keyword in male dialogue when the results are broken down by gender or genre, and so its association with male speech (made in 8.3) only surfaces in the analysis of key clusters. This supports Stubbs' (2005) argument that recurrent combinations of words add value to stylistic investigations, over and above what single words show (mentioned in 2.5.4, together with other studies demonstrating the benefits of investigating word combinations in literary texts, e.g. Mahlberg 2007, and Fischer-Starcke 2009, 2010). It also supports my decision to use a range of different methods to investigate language style in EModE plays, rather than just one process.

The examination of key results has added some general information to support the idea that Shakespeare's plays feature more deferential language than that of his contemporaries. Although the key results highlight differences, at the whole-corpus level these mainly indicate the greater extent to which Shakespeare uses language features which are also used by other dramatists, rather than language features that Shakespeare uses which his peers do not. The key cluster results provide glimpses into

ways in which Shakespeare's language appears to be distinctive, for example in his relatively greater use of *what is the matter* to elicit information, and his relatively greater use of *beseech* and *I pray you* to exaggerate the humility of requests in male and female dialogue, respectively. The analyses have not been fully conclusive, but they provide solid foundations from which to launch more detailed studies of pragmatics, particularly into the speech act of requesting and into the kinds of language used in address. Using the keyness method on smaller sections of the corpora for more focused studies would be rewarding, for example in comparisons of variation according to gender and/or social rank of characters in the plays. These variables are highly likely to be associated with deference, which has emerged as part of Shakespeare's authorial style in this chapter and in chapter 6.

The key semantic domains suggest that Shakespeare uses some concepts more than other contemporaneous dramatists (such as those surrounding the body, animals and nature), but it would not be possible to make any firm claims without evaluating the relative frequencies of individual concepts and then investigating the ways they are used in each corpus. This is outside the scope of the present study, but the most frequent concepts which populate the key domains in this study, such as nature, provide a useful place to begin a detailed follow-up analysis of the kind carried out for the single concept of love in Shakespeare's plays by Archer et al. (2009).

Although a high minimum frequency and low p value were used to help minimise the number of topical results (explained in 3.4.2), the number of stylistically useful items was fewer among the key results than among the locked results in the previous chapter. However, the other side of this is that the key results pinpointed more compatibility issues than the locked results, by virtue of their orientation to contrasts (as argued in 7.5). This has provided a useful service to the whole study,

because, as argued in the previous section, issues of non-standardisation are an inevitable limitation of working with historical source texts. Finding out what they have most impact upon is therefore sensible, in order to know the corpora better and to determine where to devote the main energies of closer analysis.

8.5.4 Evaluating the semantic domain method

As indicated in 6.1, in this section I summarise the usefulness of the semantic domains that are most frequent, locked and key, from this and the previous two chapters. In all the semantic domain categories, a proportion of the contents were incorrectly classified by the *Wmatrix* USAS tool, despite the use of the EModE tagger (mentioned in 3.3.1). This has been difficult to evaluate fully because of the need to export the *Wmatrix* concordance data for every domain and check the kinds of word forms which have been allocated a particular tag. Walker's (2012) method involving *AntConc* does this satisfactorily, but is time consuming. It is clear that the EModE dictionary of the USAS tool would need to be refined and extended to achieve greater accuracy in a more detailed study, to increase reliability and reduce the need for re-classification (as also indicated by Archer et al. 2009 and Culpeper 2009, 2011). With corpora of the size of the *SDC* and the *NDC*, this would have been an insurmountably large task in this study, however.

Since my analyses have been based on relatively large numbers of results, the classifications have been a fair guide, despite a few mis-classified items (such as *rime*, mentioned in 7.4, and some instances of *doubt*, in 6.4). Restricting the analyses to high-frequency results only, and focusing in detail only upon the domains in which the majority of words are accurately classified, has yielded some useful evidence from which to compare and contrast the most statistically frequent concepts in the dialogue

used by Shakespeare and other contemporaneous playwrights. In particular, some of the domains which lock across both corpora (in 7.4) were unexpected and interesting, such as the references to weather in EModE plays, as well as to everyday objects or aspects of social life such as personal and household items (many of which were used in metaphorical contexts). The locked domain containing language associated with "Dislike" highlighted the way in which extreme emotions and reactions are described by characters, to create the sense of drama which makes the plays more interesting and engaging for the audience. The key domain results were more topical, whereas the locked semantic domains capture the strength of feeling in the responses of characters to the topical events (such as power relations and the activities of those in power, e.g. royal or noble families). The locked domains provided a big overall picture of what is talked about in the corpora, whereas the key domains provided a more focused view of the Shakespearean drama.

The analysis of the "Clothing and personal belongings domain", which locks across both corpora (in 7.4), illustrated that semantic domains usefully capture some low-frequency results, which, taken together, represent concepts that are particularly important in the register of EModE drama. It also yielded some individual results that might not surface through other analytical processes, for example that Shakespeare creatively uses the word *cheveril* to imply flexibility, but there is no evidence of its use in the plays by the other contemporaneous dramatists in the *NDC*. The time taken to investigate disparate low-frequency results is high, though, and the benefit of doing so needs to be carefully considered in terms of the value it is likely to add; it is important to be selective about which results to follow up. The analysis of the "Calm" semantic domain, which was positively key in Shakespeare's plays, showed that grouping words by semantic association is not always a useful approach for stylistic analysis. Words

such as *gentle* and *peace* have functions which would make for more fruitful analysis in a framework geared to pragmatic analysis (through their respective roles in vocatives and attention signals).

This concludes my analyses of the quantitative data from the corpora, and I now draw my study to a close in the final chapter.

CHAPTER 9. SUMMARY AND CONCLUSIONS

9.1 Introduction

In this study, I set out to extend existing corpus stylistic research into Shakespeare's plays by examining Shakespeare's language style on an empirical basis, in the context of other plays of the period. I focused not only on differences between Shakespeare's style and that of a group of his peers, but also on similarities between them. In so doing, I tried out some new methods and built a new corpus of EModE plays as a parallel reference corpus for Shakespeare's First Folio. In 9.2 below, I briefly review the main findings from the project, and what they contribute to knowledge about Shakespeare's language style and to corpus linguistic methodology. In 9.3 I assess how well my research aims and objectives have been met, and lastly, in 9.4, I draw together some suggestions for future research.

9.2 Summary of main findings

I summarise the stylistic findings of the study in 9.2.1, and the methodological findings in 9.2.2.

9.2.1 Shakespeare's authorial style and the register of Early Modern English drama

From my analyses in chapters 6, 7 and 8, the main evidence of distinction in Shakespeare's authorial style, compared to that of the other dramatists of the Early Modern period whose work is represented in the *NDC*, is in the statistically greater extent to which deferential language is used by Shakespeare's characters. This was shown initially through the high-frequency words and 3-word clusters in 6.2 and 6.3, more of which had deferential associations in Shakespeare's plays than in the other contemporaneous plays. Further detail was added by the key results, in chapter 8, in

which deferential language associated with the mitigation of requests was shown to be positively key in Shakespeare's plays: through the keyword BESEECH (8.2), and the word clusters in which it occurs (8.3), as well as the key cluster I PRAY YOU, also in 8.2. My qualitative analysis of *beseech* indicates that Shakespeare's characters use this relatively extreme form of deference in requesting in fairly ordinary circumstances. The analysis of *I pray you* suggests that Shakespeare also opts to include a "support move" relatively more often than the other contemporaneous dramatists (i.e. that he adds it where it is not a social requirement, cf. Culpeper and Archer 2008:74). These strategies would exaggerate the emotion involved in requesting, making relatively ordinary requests seem more dramatic.

My findings also indicate that men and women in Shakespeare's plays use deferential language to a greater extent than those in the other contemporaneous plays, although in different ways: *beseech* clusters are a style feature of male Shakespearean dialogue, whereas *I pray you* is a style feature of female dialogue, particularly comedy dialogue. It seems likely that the hyperbolic emotion injected into requests made by male and female characters would have been deliberate, to add entertainment value, although it is difficult to say how far Shakespeare's audiences would have been aware of it. In 8.3 I mentioned the possibility that Shakespeare's distinctive preference for *I pray you* indicates that his authorial style is more conservative than those of his peers, and that he favours older language forms, which would support the findings of Craig (2012) and Ingram and Ingram (2012). However, the clear association with comedy dialogue and female dialogue in my data lends more support to the conclusion that Shakespeare favours it as a dramatic device, not as an older style.

Despite the distinctions noted above, the analyses of high-frequency and locked results, in chapters 6 and 7, respectively, show that Shakespeare's language

style and those of his contemporaries overlap in a great many ways. For example, Shakespeare is not unusual in favouring first-person pronouns and other function words, as well as the negative particle *not*, all of which contribute to an "interactive" and "involved" style that is typical of spoken language (Biber 1988:21, 56-58, 245; in 6.2-6.3). Plays by Shakespeare and his peers are characterised by the verbs *see*, *know*, *say*, *be*, *do*, *will* and *have*, which are frequently the central element of 3-word clusters (in 6.3 and 7.3). These serve as formulaic building blocks of dramatic dialogue which efficiently transmit characters' feelings, motivations and the essential background to the plot, which the audience needs in order to understand what is going on. Shakespeare and his peers also share some common ways of managing the communication of the play to the audience, including the certainty marker *I am sure* (the most locked 3-word cluster, in 7.3), along with relatively frequent references to concepts of everyday life (7.4).

These strategies not only convey the local on-stage actions and events of the play, but also intensify the emotion surrounding them, with metaphors and personification through pathetic fallacy (7.4). Shakespeare and the other contemporaneous dramatists also make use of frequently-occurring stagecraft devices, shown in the similar high frequency of the departure formula *farewell* (7.2), and the verbs *come* and *let* (6.2). These similarities are associated more with the register of EModE drama than with authorial styles, as argued in 7.5.1, because they demonstrably fulfil the functions of transmitting the text as a play (i.e. a dramatised story told mainly through interactional dialogue), rather than as some other kind of textual construct. The data from the investigation of the locked semantic domain "Dislike" (7.4) shows that Shakespeare and his peers made use of the emotional leverage afforded by the frequent expression of language associated with hatred and

other feelings of extreme dislike in characters' dialogue. This heightens the sense of drama. Language associated with dislike can also be instrumental in setting up events which help move the plot into a state of conflict, but this was only really clear in one play (*The Death of Robert, Earl of Huntingdon*; in example (23); see 7.4).

My findings broadly support the argument that Shakespeare uses similar vocabulary to his peers (from Craig 2011 and other scholars discussed in 2.4.3), but that he deploys this vocabulary in some relatively unusual ways and combinations (claimed by Craig 2012:6 and Crystal 2008:173, 232-233; see 3.2.2). My analyses in this study show that this kind of linguistic creativity can be explored by taking empirical language data identified at the lexical level as a starting point, then investigating its pragmatic and discoursal implications through closer examination of the surrounding co-text and context. Overall, the distinctions in Shakespeare's language style which have surfaced in my study are rooted in the **extent** to which he uses certain kinds of language, such as the self-humbling support moves in requesting, rather than that he uses language features which other dramatists do not.

Although my findings illuminate some authorial style features and some register features, they also indicate that a distinction between them is not always clear, as anticipated in 2.2. Dramatic dialogue is scripted, and, as demonstrated in my analyses in chapters 6 to 8, it contains layers of linguistic information which work simultaneously on multiple discourse levels (using Short's 1996:169-172 concept of discourse architecture, explained in 2.2). Language choices in the construction of dramatic dialogue can therefore sometimes be both functional and aesthetic, since they serve the purpose of communicating the structure, background, action and personalities in the play, as well as having artistic value. For example, it is relatively straightforward to argue that the high-frequency verb *come* (6.2) or the parting

formula *farewell* (7.2, mentioned above) are used purely for functional reasons. These language features enable the dramatist to move on-stage characters physically (Herman 1995:159-162; in 6.2), and have little or no aesthetic value. However, it is much harder to say whether the relatively frequent use of *fellow* (the most strongly locked word across both corpora) is a register feature or an authorial style feature. It is arguably a combination of both. As posited in 7.2, the use of *fellow* can have social implications, depending on the social ranks of speaker and addressee, so it functions as a marker of characters' social relations and attitudes. Signalling these to the audience is essential in playwriting, so that the audience can understand what motivates the characters' behaviour. On this basis, the pervasiveness of *fellow* in both corpora suggests it constitutes a register feature of EModE drama. Yet *fellow* is widely used outside the register of drama (judging from the *LEME* data), so it may instead be simply an artistic choice for a term of reference or address which is often made by Shakespeare and by other dramatists of his day. In this view, *fellow* constitutes a preference for a language style feature, shared by Shakespeare and other playwrights.

It is reasonable to think that a style preference which begins as an artistic choice on the part of one or two dramatists, which is then taken up by others and becomes popular, might eventually become a language feature associated with drama more than with individual dramatist(s). In other words, a style feature can evolve into a register feature, supporting the argument for seeing them on a cline rather than as discrete categories (in 2.2). A diachronic investigation of *fellow* would be a useful way of investigating this further, as I suggest in 9.4. It is also fair to conclude that the more successful EModE playwrights, such as those included in this study, are likely to have been skilled in selecting language features which are multi-functional: which they considered to have artistic value, but which also usefully contribute to

characterisation, social relationships between characters, and to the communication of the play through the characters' dialogue.

Examining Shakespeare's language style against a backdrop of other contemporaneous dramatists' styles in this study has, as indicated above, enabled me to discover some ways in which Shakespeare's use of language is exceptional, in relation to some of the language norms of the community of playwrights of which he was a part. My decision to investigate similarities as well as differences between language styles in the corpora has served the principle of giving a balanced view, and has safeguarded against the risk of exaggerating Shakespeare's exceptionality by viewing his language in isolation (a point raised in 2.6.1). More than that, investigating similarities has yielded stylistic outcomes which are of interest and value, indicating that it is a worthwhile exercise in its own right. This brings me to a summary of the findings regarding the methods used in my study.

9.2.2 Corpus linguistic methods applied in the study

I have investigated EModE plays using three types of corpus linguistic method (frequency, keyness and locking), applied to three types of language construct (words, word clusters and semantic domains), in order to base my findings about Shakespeare's style on multiple dimensions of his language (as explained in chapters 2 and 3). The high-frequency words and clusters (in 6.2 and 6.3, respectively) provided some initial indicators of potential authorial style features in Shakespeare's plays, through more high-frequency words with deferential associations. They also provide further evidence from a larger bank of data to support Culpeper and Kytö's (2010) findings, for example that EModE plays are characterised by a relatively verbal style, with many 3-word clusters which constitute utterance launchers (Biber et al.

1999:1073; see 3.2 and 6.3). However, the prevalence of utterance launchers and function words (see 6.2) found in the corpora might be expected in any corpus of spoken or speech-related language, whether historical or more modern, since they are features which are characteristic of interactional dialogue, and not specific to drama. Although illuminating relatively few details pertaining specifically to EModE plays in my study, the high-frequency results helped to confirm that the new and untested *NDC* built for the project bears the general hallmarks of a speech-related corpus, which is reassuring. They also led to some initial inroads into the investigation of Shakespeare's style which could then be taken further through the locked and keyness methods.

My research with lockwords and other locked items has broken new ground in producing evidence for:

- what locked results potentially contribute to stylistic research;
- the characteristics of locked results; and
- the way locked results can be identified by adjusting keyness software tools.

As stated in 2.7, the concept of lockwords is so new that it was difficult to know what to expect from them, particularly since they have not been investigated in EModE. Although some of the words, clusters and semantic domains which were most similar in frequency in my data could have been anticipated intuitively, such as the strength of words and concepts surrounding themes of power, death, and the objects and events of everyday social life (in 7.2), it was nevertheless useful to have these confirmed empirically. More importantly, the locked results also led to some findings which I did not anticipate, since (to me) they are not psychologically prominent (in Leech and Short's terms; see 2.4.1). In particular, these included the most strongly locked word *fellow* (7.2, highlighted among my findings in 9.2.1 above), and the two most strongly locked semantic domains: those containing weather concepts and concepts of

(extreme) dislike (7.4). As pointed out in 7.2, although *fellow* is mentioned in other corpus studies (see for example B. Busse 2006 and U. Busse 2002b), it has not been the focus of special attention, whereas in my study it surfaces as an item of particular interest through quantitative significance when the locking method is applied. The same is true of weather concepts: weather has been noted as having important meanings in Shakespeare's plays (for example by Richmond 2002; see 7.4), but has not attracted attention on a statistical basis in other corpus studies. The apparent lack of (psychological) prominence of these language features means that they are also probably unlikely to draw attention in manual stylistic analyses.

My detailed qualitative analyses of the locked results, in chapter 7, tested out whether or not they could be useful for stylistic analysis in two synchronic corpora, adding to Baker's (2011) initial findings from investigating locking across four diachronic corpora. The outcomes lead me to conclude that locked results are similar to key results and other frequency-based data, in that their statistical significance can, but does not always, point to stylistic or interpretative significance (discussed in 2.7, with reference to McIntyre 2010:168 and other corpus stylisticians). The exploration of the use of *fellow* did not lead directly to any definite conclusions about its stylistic effects in plays. However, my findings indicate that it has particular pragmatic functions when used in drama, which are perhaps linked to its pejorative implications when used in other registers (from the *LEME* data). The investigations of weather concepts showed clearer evidence of powerful stylistic effects, through their use in personification through pathetic fallacy. These would be worth pursuing in further detail, as I suggest in 9.4. The locked results have, therefore, made a valuable contribution to the study by revealing some stylistically interesting language features

which have not hitherto attracted special notice in other studies, and which have not emerged through the other methods used in this study.

My findings in chapter 7 show that lockwords fall into three identifiable kinds which have also been noted among keywords (by Scott e.g. 2000 and other scholars, in 2.7). As well as some of stylistic interest, others reflect "aboutness", although these are more generalised and thematic than the localised topical results typically found in keywords output. Also, one proper noun arose in the lockwords.

The key results in chapter 8 usefully pinpointed some potential authorial style features in Shakespeare's plays, as summarised in 9.2.1 above. This was reasonable to anticipate, given the body of existing keyness research including that into Shakespeare's plays (discussed in 2.5, e.g. Archer and Bousfield 2010; Culpeper 2002, 2009; Scott and Tribble 2006). Furthermore, as argued in 8.5, the keyness method also flagged some compatibility issues between the texts in the two corpora, which were much less clear in the output of the other methods (particularly the lockwords). On the one hand, results from the other methods were less susceptible than keywords to the inherent problems of non-standard spelling, punctuation and grammar in EModE texts, which hamper the orthographic matching processes (as explained in 5.4), so more of them were useful for analysis. On the other hand, the lack of problematic results actually masks the extent to which retrieval problems may be biasing them. Therefore, although more of the key results were discountable through problems of non-standardisation of language in the texts, this helped me to identify the strengths and weaknesses of the corpora, and served as a quality control process in the study. The outcomes of the key results make clear that although not every language feature in the output from historical corpora will be reliable enough to follow up (such as the pragmatic markers *O*, *ay* and *ha*, in 8.2), those which are reliable justify the

application of corpus methods to historical texts. In my study, the evidence that Shakespeare uses relatively more deferential language than other contemporaneous dramatists would probably have been impossible to trace through manual analysis.

The analysis of word clusters and semantic domains added value to the study by providing additional dimensions of language to that of single words, in many cases strengthening and/or enlarging the evidence from the single-word data. For example, the word clusters led to some useful pragmatic analysis, particularly in the case of the key cluster I PRAY YOU in 8.3, and the semantic domains provided a clearer picture of the "aboutness" of Shakespeare's plays compared to those of other dramatists (e.g. Shakespeare's relatively greater use of body and nature concepts, in 8.4, and the preference he shares with other dramatists for using weather concepts, noted above and discussed in 7.4).

9.2.3 Building and preparing corpora of historical texts

Building a parallel reference corpus for Shakespeare's plays in this study was a considerable undertaking, but one which provided:

(i) maximum control over the quality of the contents; and

(ii) the ability for the contents to be manipulated in order to apply the desired methods (4.3.1).

There was notable similarity between the most frequent 3-word clusters in my data and those in Culpeper's (2011:73) lexical bundle data from Shakespeare's plays using the larger and broader *KEMPE* corpus as a comparator (in 6.3). This finding surprised me, because of the difference in the size and content of the reference corpora of EModE plays used in his study and mine, but lends support to the argument that a "robust core" of (key) results can be obtained using different kinds of reference

corpora (Scott and Tribble 2006:64; see 3.6). The results I examined in this study are all high-frequency, however, generated from both corpora in their entireties. It may be that analyses of component parts of the corpora, such as one genre or a particular type of character, would be more affected by the contents of the reference corpus. Results of lower frequency would be more susceptible to bias in the contents of the reference corpus towards a particular sub-genre or authorial style (factors discussed in chapter 4). A relatively narrow, specialised parallel reference corpus such as the *NDC* would therefore appear to offer greater certainty of generating results which reflect Shakespeare's language style features, although some of these might well arise with broader reference corpora.

The lack of standardisation in spelling and punctuation in historical texts is certain to remain a thorn in the side of corpus linguists who work with them, but my study has tested out the relative benefits of older and newer versions of the spelling regularisation software *VARD 2*, which has been shown to improve the reliability of retrieval of frequency-based results (e.g. Archer et al. 2003; Rayson 2007; see 5.4). The researcher must decide where to draw the line between standardisation which is desirable to improve the potential for orthographic matching, and editorial intervention (particularly in the form of modernisation) which could rub out historical language variation that might be of interest. Where this line is drawn depends on what the researcher wishes to investigate, and this will influence the amount of regularisation that can be undertaken using automated methods. My findings in 5.4 showed that the newer version of *VARD 2* (version 2.3) is more beneficial for a study where some modernisation is desirable, since it is based on a modern dictionary which can be trained to recognise some EModE features and leave them unchanged. I showed in 5.4 that training is necessary because otherwise it will merge potentially interesting

distinctions, e.g. between *you* and *thou*. My findings support those of Baron and Rayson (2009:8-14), who argue that the benefits of training *VARD 2* decline beyond a certain amount of sample data (12,000 words in their study; 20,000 words in my study). *VARD 2* version 2.1.5 is much slower to run than V.2.3, but preserves more historical language features, since it is based on known EModE variants and not on modern dictionaries, and it may be a better choice in studies where more conservative regularisation is desirable.

There also remains the problem of not knowing how much regularisation has already been done by the compositors of early extant printed texts, as argued in 5.4 with regard to s-genitives and the apostrophe. This means that it will sometimes be impossible to judge whether or not a result is reliable (in the absence of copies of the original manuscripts), as in the case of the pragmatic markers *O* and *ha* (in 8.2). Given that discussion space is limited in any piece of research, it is necessary to be selective about which results to include, and reliability can be considered the first inclusion criterion in deciding what to present in a corpus study of historical texts. I would expect the accuracy problems identified in some of the *EEBO* texts (in 5.3) to be eliminated over time, as the re-keying process continues during the ongoing *EEBO-TCP* project (according to Hope 2011).

My findings from using the scripting language PHP (in 5.2) show that it has much potential for reducing the labour-intensity of annotation, through its ability to carry out multiple commands in multiple texts from one script. The annotation of the *NDC* was made more efficient and quick through the application of just five PHP scripts, especially the automated tagging of speaker-id labels for characters in the play-texts. It is ideal for annotating numerous items of similar form, but tagging meta-data which is localised, such as stage directions in my data, still needs to be carried out

manually. Writing PHP scripts does, however, require considerable knowledge and expertise of computer programming, which may not be feasible to acquire in the course of a project. I could not have used them without expert help. The scripts used for the annotation in this study may well be adaptable to other corpora which require similar kinds of search-and-replacement annotation with XML tags. For example, there is potential for adding meta-data such as tags for gender or social rank of characters using PHP, by linking them to speaker-id tags. The text editor used with PHP in this study, *Notepad++* offers a number of advantages that make annotation more efficient, such as editing with multiple open documents using regular expressions. As argued in 5.5, it may be a better option for researchers without the resources to acquire sufficient programming skills to use PHP for annotation.

As summarised above, my study has tested out some existing software in new ways (the adaptation of the keyness tools to obtain locked results, and the use of PHP for corpus annotation), as well as comparing the benefits of newer and previous versions of *VARD 2*, which is itself a new resource still undergoing development.

Having reviewed the main findings of my study, both stylistic and methodological, I will now assess the extent to which I have been able to answer my research questions and how well the study has achieved its aims.

9.3 Reflections on the achievement of the aims of the study

My study was structured around three main research questions, given in 1.3, in order to achieve its aims. These aims, which were stated in 1.2, set out to make a contribution to knowledge about:

(i) the extent to which Shakespeare and some of his contemporaries share preferences for certain language styles in the construction of dramatic dialogue, and in what ways Shakespeare's style appears to be distinctive;

(ii) the value of a selection of corpus linguistic methods in investigating (i) above, especially the new "locking" method; and

(iii) the construction and treatment of historical corpora.

In this section I offer some conclusions about how well each area has been addressed in my study.

My first research question concerned what quantitatively significant words, word clusters and semantic domains would reveal about Shakespeare's language style in the context of wider EModE drama. This has been answered by my analyses in chapters 6 to 8 (summarised in 9.2.1 above). The differences which mark Shakespeare's style out as distinctive or exceptional compared to other contemporaneous dramatists are subtle, but a picture has emerged of his relatively greater use of deferential language (from the single word and word cluster results) which I have argued would inject relatively more emotion into ordinary situations. Ordinary situations pervade the plays by Shakespeare and by his peers, as was shown by the number of semantic domains concerning everyday social life which lock across both corpora, in 7.4. My analyses have successfully captured some ways in which the treatment of ordinary situations is enhanced by Shakespeare with language that, while not actually different to that used by other playwrights of his day, is applied in a different way. Although this has been argued as being the crux of what makes Shakespeare's language style special and outstanding compared to other playwrights of the period (e.g. by Craig 2012 and Crystal 2008, as noted in 9.2.1), it is not actually easy to pin down empirically. The matrix of methods and language constructs used in

my study has been wide-ranging, and has used lexical, lexico-grammatical and semantic data as a starting point for drilling down to pragmatic effects, in which the distinctiveness of Shakespeare's style seems to lie. Although demonstrably worthwhile, the analytical route to uncovering them was lengthy and involved a lot of prospecting and panning to turn up some specific findings. This is, however, the first substantial corpus stylistic study which compares Shakespeare's plays to a parallel corpus of other contemporaneous plays, and my analyses break some new ground that can be explored in closer detail in future research (as I suggest in the next section).

My second research question focused on the methods for investigating similarities between Shakespeare's style and that of his peers. My analyses in chapters 7 and 8 confirm that locked results provide as much scope for useful stylistic analysis as key results. My study demonstrates that these methods can be deployed as a dual approach to investigating language in two corpora, using the same software tool, to provide a more balanced view than is afforded through key results alone (answering research question 2.1). By way of addressing research question 2.2, my experiences in this study show that the inherent retrieval problems caused by non-standard spelling, grammar and punctuation in EModE texts do not surface in the locked results output as they do with keyness output. The orientation to similarity means that the computer software simply finds and displays what matches most often, so the locked output does not contain the kinds of rogue results which arise from problems with orthographic matching (discussed in 8.2). More of the output from locked method is therefore likely to be reliable than from the keyness method when used with Early Modern texts, although the underlying issue of statistical reliability is still present: there may be results which do not surface as locked simply because they exist in variant forms in the corpora. My study demonstrates that Baker's (2011) locking concept can be

applied synchronically as well as diachronically, although as I argue in the next section, the method needs further testing.

My third and final research question considered the issues involved in building a specialised parallel reference corpus for Shakespeare's plays. In my discussions of other corpora of EModE drama, in chapter 4, I explained that on the one hand there is now a vast source of digitised contemporaneous dramatic dialogue (on *EEBO*) which other scholars have already made use of (answering research question 3.1). On the other hand, to address research question 3.2, the diversity of its content means that careful choices need to be made in order to construct a collection that is most likely to facilitate the investigation of authorial style, rather than simply showing evidence arising from variation in genre features or style change over time. As indicated in 9.2.3 above, however, my study does not show that a parallel reference corpus necessarily generates results which are more useful than those from a broad reference corpus of EModE drama such as that used by Culpeper (2011), although the *NDC* constructed for this study offers other advantages such as ease of manipulation and access to separable component parts and contents (argued in 4.4).

Research questions 3.3 and 3.4 respectively concerned the identification of ways of rendering the texts more suitable for exploitation with corpus tools, and how this can be carried out. These were answered in chapter 5, where I argued that the main considerations are: the annotation of the corpus texts, the improvement of accuracy in some of the digitised play-texts on *EEBO*, and the regularisation of historical spelling variation to improve the potential for retrieving results. I also explained how these processes were carried out. In addressing research question 3.5, in 5.2 I explained that I opted to undertake the essential bare minimum of annotation, which would exclude all the non-dialogic text from computation and add only some

basic meta-data to each file. This took several months even with the added efficiency afforded by the application of PHP scripts, but was completely achieved. In contrast, the improvement of accuracy of the texts and the regularisation of non-standard spelling, including punctuation in compounds and contractions, was addressed rather than definitively solved. As indicated in 9.3.3 above, I, like other corpus linguists working with historical texts, was faced with the paradox of improving the potential for orthographic matching at the expense of potentially erasing historical language features. I did not attempt to quantify the benefit of the spelling regularisation carried out on the corpora in my study, but I spent as much time on training the *VARD 2* software as the project reasonably allowed, in recognition of the improvement in reliability of retrieving results noted in other research (e.g. Archer et al. 2003; Lehto et al. 2010; Rayson 2007). Although a few problematic results arose from non-standard spelling and punctuation (mainly in 8.2), overall the study was not substantially affected by them, which I regard as justifying my efforts to minimise the known difficulties.

My study has achieved its main aims and thereby contributed to the field of corpus stylistics, particularly the investigation of Shakespeare's language style. There were, however, some limitations of working with historical texts, and some of my findings need exploring further to reach deeper and clearer conclusions. The potential for future research arising from my study concerns both stylistic investigation and methodological testing, as well as the further exploitation of the *NDC*, as I now discuss briefly in the final section.

9.4 Suggestions for future research

My findings surrounding the relatively more frequent use of *beseech* and *I pray you* by men and women, respectively, in Shakespeare's plays compared to other contemporaneous plays would be worth further investigation. In particular, it would be useful to analyse these results by social rank and gender of speaker and addressee, to see what kinds of characters typically use these very deferential "support moves" in making requests. The use of *fellow* as a term of reference and address merits further exploration in EModE in that it would be interesting to see how its use varies over time, and across a range of EModE text-types in addition to drama. The *CED* would be a potential resource for a sociolinguistic study of *fellow*. It would be useful to analyse the social status of speakers and addressees by whom and to whom it is used, to build up a picture of the pejorative implications of the term, and thereby to say more about the stylistic effects it has in dramatic dialogue.

Much has been written about the representation of women in EModE plays, and in 8.4 I linked some of my findings from the locked semantic domains to research by literary critical scholars (e.g. Burkert 2011, Jardine 1983:141-168 and Orgel 1996), as well as to my own previous research into gender in Shakespeare's plays (Demmen 2009) in the course of analysing the key cluster results in 8.3. A comparative corpus study focusing on gender and language styles would usefully take this much further, through a comparison of:

- the internal variation between male and female dialogue in Shakespeare's plays and between male and female dialogue in the other contemporaneous plays; and

- the external variation in language use between characters of each sex in the parallel sections of both corpora (to see how Shakespeare constructs different genders, in comparison to other dramatists of the period).

Empirical comparisons of the language styles in different dramatic genres would also be useful, to contextualise what has been said in many literary critical studies about Shakespeare's style of writing comedy, history and tragedy (by comparing it with data from other contemporaneous plays).

A diachronic study would be helpful in the case of the word *fellow*, to see if there is any evidence that it is a style feature in the work of a few dramatists which is taken up by others over time, and which eventually constitutes more of a register feature (discussed as a possibility in 9.2.2).

The phrase *I am sure*, which occurs as the most strongly locked 3-word cluster in the corpora, would be worth investigating further with regard to the social rank and gender of characters who use it most in EModE plays. This is in view of Fischer-Starcke's (2010:123) findings that it is an authorial feature of Jane Austen's novels, several centuries later, in which it has implications for characterisation as well as catching the attention of the reader (7.3).

Address forms would be a potentially rich seam to mine in the other contemporaneous drama, to make some comparisons with existing research into Shakespeare's plays (notably by B. Busse 2006 and U. Busse 2002b), as mentioned in my findings in 6.3., 7.4 and 8.4).

If the reliability of the EModE tagger in the *Wmatrix* USAS tool can be improved (discussed in 6.4, 7.4, 8.4 and 8.5.4), the other contemporaneous plays could be usefully investigated much further via key semantic domains (to see, for instance, if and how the concept of love is constructed differently by Shakespeare's

contemporaries, following the approach of Archer et al. 2009). The locked semantic domain concerning weather concepts also provides scope for further research, particularly its use in metaphor and personification through pathetic fallacy.

With regard to the methodology used in this study, there is much more work to be done to further the investigation of similarities between corpora. There is a need for more testing of corpus tools and methods for investigating language similarities, including Baker's (2011) locking concept. The ready adjustability of the keyness programmes in *WordSmith* and *Wmatrix* to the investigation of locked results will, I hope, encourage others to test out the method and investigate its implications and reliability. Envisaging a lockword as the opposite of a keyword (Baker 2011:73) is a useful place to start, and from the analyses in this study I suggest that it is also helpful to consider locked language features as evidence for shared preferences, and key language features as evidence for contrasting preferences (among the authors, speakers, writers or other originators of the texts in the corpora under investigation). Locking is dependent on high frequency and similarity of frequency, and will benefit from further research into different cut-off points, as has the keyness method (from studies such as Rayson et al. 2004).

With regard to the corpus constructed for this project, the 796,582-word *NDC* could easily be adapted from a specialised parallel reference corpus for Shakespeare's plays to become a more general corpus of EModE plays. It could then be investigated further on its own, or in comparison to corpora of other EModE registers and text-types. As noted in 1.2, drama is of historical sociolinguistic interest, as well as of (literary) stylistic interest. This is demonstrated by Culpeper and Kytö (2010), and also shown briefly in my discussion of the lockword *farewell* in 7.2, which could be linked to Arnovick's (1999:96) findings. The *NDC* would need to be enlarged slightly to be

more representative of the register of EModE drama, mainly by including a selection of Shakespeare's plays, and by adding a few more city comedies, because these are under-represented due to not being a typical type of Shakespeare's comedy (explained in 4.3.2.2). Dekker's *The Shoemaker's Holiday* (dated 1599) and Middleton's *The Roaring Girl* (1611) would be potential additions from the existing date band of the *NDC*. The date band could also be widened to include plays from earlier or later years. For example, extending the corpus by adding plays from later years would allow the inclusion of some comedies of manners, a popular genre which developed from the city comedy tradition. Early extant versions of play-texts from *EEBO*, with annotation and spelling regularisation as detailed in chapter 5, would be preferable as source texts for Shakespeare's plays if the *NDC* is enlarged, not the modernised play-texts in the *SDC*, to minimise the textual compatibility issues.

Finally, and in the longer term, a comparative corpus-based encyclopaedia or dictionary of Shakespeare's plays, such as is proposed by Culpeper (2011), would be a valuable resource for development. This would provide much-needed support to researchers in both linguistic and literary disciplines who work with EModE drama. Some justification for this was demonstrated in 7.2, in that the route to investigating the meaning of *fellow* via the *OED* proved to be somewhat circular, because most of the attestations of its use in the Early Modern period are from Shakespeare's works. As argued at the start of this study, the unique place accorded to Shakespeare's plays, and the claims of exceptionality of his language, need to be viewed in the wider context of other plays of the period, on an empirical basis. This is not to detract from them in any way, but rather to enrich knowledge about them and thereby reach a more profound understanding of why they are of lasting interest, popularity and critical acclaim.

REFERENCES

Primary sources

Facsimile printed manuscripts:

The Battle of Alcazar 1594. Edited by Greg, W.W., 1907. Chiswick: Chiswick Press; Oxford: Oxford University Press.

The Valiant Welshman (1615). Edited by Farmer, J.S., 1913. Amersham: Issued for The Tudor Facsimile Texts.

Mike Scott's Shakespeare Corpus:

Shakespeare Corpus. Downloaded 2007. See http://www.lexically.net/wordsmith/support/shakespeare.html (last accessed 10.08.12).

Plays downloaded October 2010 from *Early English Books Online, 1475-1700*. ProQuest LLC. See http://eebo.chadwyck.com (last accessed 11.08.12):

A moste excellent comedie of Alexander, Campaspe, and Diogenes played beefore the Queenes Maiestie on twelfe day at night, by her Maiesties children, and the children of Poules. , Imprinted at London : [By Thomas Dawson] for Thomas Cadman, 1584. STC (2nd ed.) / 17047.5.

A most pleasant comedie of Mucedorus the kings sonne of Valentia and Amadine the Kings daughter of Arragon with the merie conceites of Mouse. Newly set foorth, as it hath bin sundrie times plaide in the honorable cittie of London. Very delectable and full of mirth. , London : Printed for William Iones, dwelling at Holborne conduit, at the signe of the Gunne, 1598. STC (2nd ed.) / 18230.

A pleasant comedy entituled: An humerous dayes myrth As it hath beene sundrie times publikely acted by the right honourable the Earle of Nottingham Lord high Admirall his seruants. By. G.C. , At London : Printed by Valentine Syms, 1599. STC (2nd ed.) / 4987.

A pleasant conceited comedie, wherein is shewed, how a man may chuse a good wife from a bad As it hath bene sundry times acted by the Earle of Worcesters seruants. , London : Printed [by T. Creede] for Mathew Lawe, and are to be solde at his shop in Paules Church-yard, neere vnto S. Augustines gate, at the signe of the Foxe, 1602. STC (2nd ed.) / 5594.

A woman kilde with kindnesse. Written by Tho. Heywood , London : Printed by William Iaggard dwelling in Barbican, and are to be sold in Pauls Church-yard. by Iohn Hodges, 1607. STC (2nd ed.) / 13371.

Bartholmew fayre : a comedie, acted in the yeare, 1614 by the Lady Elizabeths seruants, and then dedicated to King Iames, of most blessed memorie ; The diuell is an asse : a comedie acted in the yeare, 1616, by His Maiesties seruants ; The staple of newes : a comedie acted in the yeare, 1625, by His Maiesties seruants by the author,

Beniamin Iohnson. , London : Printed by I.B. for Robert Allot, and are to be sold at the signe of the Beare, in Pauls Church-yard, 1631. STC (2nd ed.) / 14753.5.

Gallathea As it was playde before the Queenes Maiestie at Greene-wiche, on Newyeeres day at night. By the Chyldren of Paules. , At London : Printed by Iohn Charlwoode for the vviddow Broome, 1592. STC (2nd ed.) / 17080.

If you knovv not me, you know no bodie: or, The troubles of Queene Elizabeth , At London : Printed [by Thomas Purfoot] for Nathaniel Butter, 1605. STC (2nd ed.) / 13328.

Mr. VVilliam Shakespeares comedies, histories, & tragedies Published according to the true originall copies. , London: Printed by Isaac Iaggard, and Ed. Blount [at the charges of W. Iaggard, Ed. Blount, I. Smithweeke, and W. Aspley], 1623. STC (2nd ed.) / 22273.

Tamburlaine the Great Who, from a Scythian shephearde, by his rare and woonderfull conquests, became a most puissant and mightye monarque. And (for his tyranny, and terrour in warre) was tearmed, the scourge of God. Deuided into two tragicall discourses, as they were sundrie times shewed vpon stages in the citie of London. By the right honorable the Lord Admyrall, his seruauntes. , London : Printed by Richard Ihones: at the signe of the Rose and Crowne neere Holborne Bridge, 1590. STC (2nd ed.) / 17425.

The battell of Alcazar fought in Barbarie, betweene Sebastian king of Portugall, and Abdelmelec king of Marocco. With the death of Captaine Stukeley. As it was sundrie times plaid by the Lord high Admirall his seruants [by George Peele]. , Imprinted at London : By Edward Allde for Richard Bankworth, and are to be solde at his shoppe in Pouls Churchyard at the signe of the Sunne, 1594. STC (2nd ed.) / 19531.

The blinde begger of Alexandria most pleasantly discoursing his variable humours in disguised shapes full of conceite and pleasure. As it hath been sundry times publickly acted in London. by the right honorable the Earle of Nottingham, Lord high Admirall his seruantes. By George Chapman: Gentleman. , Imprinted at London : [By J. Roberts] for William Iones, dwelling at the signe of the Gun, neere Holburne Conduict, 1598. STC (2nd ed.) / 4965.

The bond-man an antient storie. As it hath been often acted with good allowance, at the Cock-pit in Drury-lane: by the most excellent princesse, the Lady Elizabeth her Seruants. By Phillip Massinger. , London : Printed by Edw: Allde, for Iohn Harison and Edward Blackmore, and are to be sold at the great south dore of Pauls, 1624. STC (2nd ed.) / 17632.

The changeling as it was acted (with great applause) at the Privat house in Drury-Lane, and Salisbury Court / written by Thomas Midleton, and William Rowley, Gent. , London : Printed for Humphrey Moseley ..., 1653. Wing / M1980.

The death of Robert, Earle of Huntington Otherwise called Robin Hood of merrie Sherwodde: with the lamentable tragedie of chaste Matilda, his faire maid Marian, poysoned at Dunmowe by King Iohn. Acted by the Right Honourable, the Earle of

Notingham, Lord high Admirall of England, his seruants. , Imprinted at London : [By R. Bradock] for William Leake, 1601. STC (2nd ed.) / 18269.

The fair maid of the West. Or, A girle worth gold. The first part. As it was lately acted before the King and Queen, with approved liking. By the Queens Majesties Comedians. Written by T.H. , London : Printed [by Miles Flesher] for Richard Royston, and are to be sold at his shop in Ivie Lane, 1631. STC (2nd ed.) / 13320.

The faithfull shepheardesse. By Iohn Fletcher. , Printed at London : [By Edward Allde] for R. Bonian and H. Walley, and are to be sold at the spred Eagle ouer against the great north dore of S. Paules, [1610?]. STC (2nd ed.) / 11068.

The famous chronicle of king Edward the first, sirnamed Edward Longshankes with his returne from the holy land. Also the life of Lleuellen rebell in Wales. Lastly, the sinking of Queene Elinor, who sunck at Charingcrosse, and rose againe at Pottershith, now named Queenehith. , London : Printed by Abell Ieffes, and are to be solde by William Barley, at his shop in Gratious streete, 1593. STC (2nd ed.) / 19535.

The famous history of Sir Thomas VVyat With the coronation of Queen Mary, and the coming in of King Philip. As it was plaied by the Queens Maiesties Seruants. Written by Thomas Dickers, and Iohn Webster. , London : Printed by E[dward] A[llde] for Thomas Archer, and are to be solde at his shop in the Popes-head Pallace nere the Royall Exchange, 1607. STC (2nd ed.) / 6537.

The famous tragedy of the rich Ievv of Malta As it vvas playd before the King and Queene, in his Majesties theatre at White-hall, by her Majesties Servants at the Cock-pit. Written by Christopher Marlo. , London : Printed by I[ohn] B[eale] for Nicholas Vavasour, and are to be sold at his shop in the Inner-Temple, neere the Church, 1633. STC (2nd ed.) / 17412.

The first and second partes of King Edward the Fourth Containing his mery pastime with the tanner of Tamworth, as also his loue to faire Mistrisse Shoare, her great promotion, fall and miserie, and lastly the lamentable death of both her and her husband. Likewise the besieging of London, by the bastard Falconbridge, and the valiant defence of the same by the Lord Maior and the citizens. As it hath diuers times beene publikely played by the Right Honorable the Earle of Derbie his seruants. , Imprinted at London : By F[elix] K[ingston] for Humfrey Lownes and Iohn Oxenbridge, 1600. STC (2nd ed.) / 13342.

The first part of the true and honorable historie, of the life of Sir Iohn Old-castle, the good Lord Cobham As it hath been lately acted by the right honorable the Earle of Notingham Lord high Admirall of England his seruants. , London : Printed by V[alentine] S[immes] for Thomas Pauier, and are to be solde at his shop at the signe of the Catte and Parrots neere the Exchange, 1600. STC (2nd ed.) / 18795.

The honorable historie of frier Bacon, and frier Bongay As it was plaid by her Maiesties seruants. Made by Robert Greene Master of Arts. , London : Printed [by Adam Islip] for Edward White, and are to be sold at his shop, at the little north dore of Poules, at the signe of the Gun, 1594. STC (2nd ed.) / 12267.

The lamentable and true tragedie of M. Arden of Feuersham in Kent Who was most wickedlye murdered, by the meanes of his disloyall and wanton wyfe, who for the loue she bare to one Mosbie, hyred two desperat ruffins Blackwill and Shakbag, to kill him. VVherin is shewed the great malice and discimulation of a wicked woman, the vnsatiable desire of filthie lust and the shamefull end of all murderers. Imprinted at London : [By E. Allde] for Edward White, dwelling at the lyttle north dore of Paules Church at the signe of the Gun, 1592. STC (2nd ed.) / 733.

The life of the dutches of Suffolke As it hath beene divers and sundry times acted, with good applause. , [London] : Imprinted by A. M[athewes] for Iasper Emery; at the Flowerdeluce in Paules Church-yard, 1631. STC (2nd ed.) / 7242.

The maides tragedy As it hath beene diuers times acted at the Blacke-friers by the Kings Maiesties Seruants. , London : Printed [by Nicholas Okes] for Richard Higgenbotham and are to be sold at the Angell in Pauls Church-yard, 1619. STC (2nd ed.) / 1677.

The massacre at Paris with the death of the Duke of Guise. As it was plaide by the right honourable the Lord high Admirall his Seruants. Written by Christopher Marlow. , At London : Printed by E[dward] A[llde] for Edward White, dwelling neere the little north doore of S. Paules Church, at the signe of the Gun, [1594?]. STC (2nd ed.) / STC (2nd ed.) / 17423.

The miseries of inforst mariage As it is now playd by his Maiesties Seruants. By George Wilkins. , London : Printed [by William Jaggard] for George Vincent, and are to be sold at his shop in Woodstreet, 1607. STC (2nd ed.) / 25635.

The pleasant comedie of old Fortunatus As it was plaied before the Queenes Maiestie this Christmas, by the Right Honourable the Earle of Nottingham, Lord high Admirall of England his seruants. , London : Printed by S. S[tafford] for William Aspley, dwelling in Paules Church-yard at the signe of the Tygers head, 1600. STC (2nd ed.) / 6517.

The pleasant history of the two angry women of Abington With the humorous mirth of Dicke Coomes and Nicholas Prouerbes, tvvo seruingmen. As it was lately playde by the right Honorable the Earle of Nottinghamn, Lord high Admirall his seruants. By Henry Porter Gent. , Imprinted at London : [By Edward Allde] for VVilliam Ferhrand [i.e. Ferbrand], and are to be solde at his shop at the corner of Colman streete neere Loathbury, 1599. STC (2nd ed.) / 20122.

The old wiues tale A pleasant conceited comedie, played by the Queenes Maiesties players. Written by G.P. , Printed at London : By Iohn Danter, and are to be sold by Raph Hancocke, and Iohn Hardie, 1595. STC (2nd ed.) / 19545.

The Scottish historie of Iames the fourth, slaine at Flodden Entermixed with a pleasant comedie, presented by Oboram King of Fayeries: as it hath bene sundrie times publikely plaide. Written by Robert Greene, Maister of Arts. , London : Printed by Thomas Creede, 1598. STC (2nd ed.) / 12308.

The Spanish tragedie containing the lamentable end of Don Horatio, and Bel-imperia: with the pittifull death of olde Hieronimo. , At London : Printed by Edward Allde, for Edward White, [1592] STC (2nd ed.) / 15086.

The tragedy of the Dutchesse of Malfy As it was presented priuatly, at the Black-Friers; and publiquely at the Globe, by the Kings Maiesties Seruants. The perfect and exact coppy, with diuerse things printed, that the length of the play would not beare in the presentment. VVritten by Iohn Webster. , London : Printed by Nicholas Okes, for Iohn Waterson, and are to be sold at the signe of the Crowne, in Paules Church-yard, 1623. STC (2nd ed.) / 25176.

The tragedie of Dido Queene of Carthage played by the Children of her Maiesties Chappell. Written by Christopher Marlowe, and Thomas Nash. Gent. Actors Iupiter. Ganimed. Venus. Cupid. Iuno. Mercurie, or Hermes. AEneas. Ascanius. Dido. Anna. Achates. Ilioneus. Iarbas. Cloanthes. Sergestus. At London : Printed, by the widdowe Orwin, for Thomas Woodcocke, and are to be solde at his shop, in Paules Church-yeard, at the signe of the blacke Beare, 1594. STC (2nd ed.) / 17441.

The tragicall history of D. Faustus As it hath bene acted by the right honorable the Earle of Nottingham his seruants. Written by Ch. Marl. , London : Printed by V. S[immes] for Thomas Bushell, 1604. STC (2nd ed.) / 17429.

The troublesome raigne and lamentable death of Edward the second, King of England with the tragicall fall of proud Mortimer: as it was sundrie times publiquely acted in the honourable citie of London, by the right honourable the Earle of Pembrooke his seruants. Written by Chri. Marlow Gent. Imprinted at London : [By R. Robinson] for William Iones dwelling neere Holbourne conduit, at the signe of the Gunne, 1594. STC (2nd ed.) / 17437.

The valiant VVelshman, or The true chronicle history of the life and valiant deedes of Caradoc the Great, King of Cambria, now called Wales As it hath beene sundry times acted by the Prince of Wales his seruants. Written by R.A. Gent. , London : Imprinted by George Purslowe for Robert Lownes, and are to be solde at his shoppe at the little north dore of Paules, 1615. STC (2nd ed.) / 16.

The white diuel, or, The tragedy of Paulo Giordano Vrsini, Duke of Brachiano with the life and death of Vittoria Corombona the famous Venetian curtizan. Acted by the Queenes Maiesties Seruants. Written by Iohn Webster. , London : Printed by N[icholas] O[kes] for Thomas Archer, and are to be sold at his shop in Popes head Pallace, neere the Royall Exchange, 1612. STC (2nd ed.) / 25178.

The vvoman hater As it hath beene lately acted by the Children of Paules [author Francis Beaumont]. , London : Printed [by Robert Raworth], and are to be sold by Iohn Hodgets in Paules Church-yard, 1607. STC (2nd ed.) / 1693.

The workes of Beniamin Ionson. , London : Printed by W: Stansby, and are to be sould by Rich: Meighen, An0 D. 1616. STC (2nd ed.) / 14752.

Two new playes ... written by Tho. Middleton, Gent. , London : Printed for Humphrey Moseley, and are to be sold at his shop ..., 1657. Wing / M1989.

Secondary sources

Adolphs, S. (2006) *Introducing Electronic Text Analysis. A Practical Guide for Language and Literary Studies.* Abingdon and New York: Routledge.

Aijmer, K. (1996) *Conversational Routines in English.* London and New York: Addison Wesley Longman.

Akimoto, M. (2000) "The grammaticalization of the verb 'pray'". In Fischer, O, Rosenbach, A., Stein, D. (eds.) *Pathways of Change. Grammaticalization in English.* Amsterdam: John Benjamins, pp. 67-84.

Allan, K. (2010) "Tracing metonymic polysemy through time: MATERIAL FOR OBJECT mappings in the *OED*". In Winters, E., Tissari, H. and Allan, K. (eds.) *Historical Cognitive Linguistics*, pp. 163-96.

Alexander, C.M.S. (2004) (ed.) *Shakespeare and Language.* Cambridge: Cambridge University Press.

Altenberg, B. (1998) "On the phraseology of spoken English: The evidence of recurrent word-combinations". In Cowie, A.P. (ed.) *Phraseology: Theory, Analysis and Applications.* Oxford: Clarendon Press, pp. 101-22.

Anthony, L. (2007) *AntConc* (Version 3.2.1w). Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/ (accessed 20.02.12).

Archer, D. (2007) "Computer-assisted literary stylistics: The state of the field". In Lambrou, M. and Stockwell, P. (eds.), pp. 244-56.

Archer, D. (ed.) (2009) *What's in a Word-List? Investigating Word Frequency and Keyword Extraction.* Farnham, U.K. and Burlington, U.S.A.: Ashgate.

Archer, D. (2009) "Does frequency really matter?" In Archer, D. (ed.), pp. 1-15.

Archer, D. and Bousfield, D. (2010) "'See better, Lear'? See Lear better! A corpus-based pragma-stylistic investigation of Shakespeare's *King Lear*". In McIntyre, D. and Busse, B. (eds.), pp. 183-203.

Archer, D. and Culpeper, J. (2003) "Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics". In Wilson, A., Rayson, P. and McEnery, T. (eds.) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech.* Frankfurt am Main: Peter Lang, pp. 37-58.

Archer, D. and Rayson, P. (2004) "Using an historical semantic tagger as a diagnostic tool for variation in spelling". Paper given at *Thirteenth International Conference on English Historical Linguistics (ICEHL 13)*, University of Vienna, Austria, 23-29 August 2004.

Archer, D., Culpeper, J. and Rayson, P. (2009) "Love – 'a familiar or a devil'? An exploration of key domains in Shakespeare's comedies and tragedies". In Archer, D. (ed.), pp. 137-57.

Archer, D., McEnery, T., Rayson, P. and Hardie, A. (2003) "Developing an automated semantic analysis system for Early Modern English". In Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*, UCREL Technical Paper Number 16. Lancaster: UCREL, pp. 22-31.

Argamon, S., Chase, P., Hota, S.R., Garg, N., Levitan, S. and Whitelaw, C. (2007) "Stylistic text classification using functional lexical features". In *Journal of the American Society for Information Science and Technology* 58(6), 802-22.

Arnovick, L.K. (1999) *Diachronic Pragmatics. Seven Case Studies in English Illocutionary Development.* Amsterdam and Philadelphia: John Benjamins.

Aston, G. and Burnard, L. (1998) *The BNC Handbook.* Edinburgh: Edinburgh University Press.

Aston, E. and Savona, G. (1991) *Theatre as Sign-System: A Semiotics of Text and Performance.* London: Routledge.

Baker, P. (2004) "Querying keywords. Questions of difference, frequency, and sense in keywords analysis". In *Journal of English Linguistics* 32(4), 346-59.

Baker, P. (2006) *Using Corpora in Discourse Analysis.* London and New York: Continuum.

Baker, P. (2009) "The question is, how cruel is it? Keywords, fox hunting and the House of Commons". In Archer, D. (ed.), pp. 125-36.

Baker, P. (2011) "Times may change, but we will always have *money*: Diachronic variation in recent British English". In *Journal of English Linguistics* 39(1), 65-88.

Baker, P., Hardie, A. and McEnery, T. (2006) *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Balossi, G. (2009) *Language and characterization in Virginia Woolf's* The Waves*: A computer-aided approach.* Unpublished PhD thesis, Lancaster University, U.K.

Barber, C. (1981) "'You' and 'thou' in Shakespeare's *Richard III*". In *Leeds Studies in English* [New Series 12], pp. 273-89.

Barcelona Sánchez, A. (1995) "Metaphorical models of romantic love in *Romeo and Juliet*". In *Journal of Pragmatics* 24, 667-88.

Baron, A. and Rayson, P. (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.

Baron, A. and Rayson, P. (2009) "Automatic standardisation of texts containing spelling variation. How much training data do you need?" In Mahlberg, M, González-Díaz, V. and Smith, C. (eds.) *Proceedings of the Corpus Linguistics Conference, CL2009*, University of Liverpool, U.K., 20-23 July 2009.

Baron, A., Rayson, P. and Archer, D. (2009) "Word frequency and key word statistics in corpus linguistics". In *Anglistik: International Journal of English Studies* 20(1), 41-67.

Benson, P. (2001) *Ethnocentrism and the English Dictionary*. London: Routledge.

Berber Sardina, T. (1999) "Using keywords in text analysis: Practical aspects". In *DIRECT Working Papers* 42, São Paulo and Liverpool. See http://www2.lael.pucsp.br/direct/DirectPapers42.pdf (last accessed 10.08.12).

Berry, D.A. (1996) *Statistics: A Bayesian Perspective.* Belmont, California: Wadsworth.

Berry, H. (2002) "Playhouses". In Kinney, A.F. (ed.), pp. 148-62.

Biber, D. (1988) *Variation Across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (1993) "Representativeness in corpus design". In *Literary and Linguistic Computing* 8(4), 243-57.

Biber, D. (2009) "A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and writing". In *International Journal of Corpus Linguistics* 14(3), 275-311.

Biber, D. and Barbieri, F. (2007) "Lexical bundles in university spoken and written registers". In *English for Specific Purposes* 26(3), 263-86.

Biber, D. and Conrad, S. (2009) *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Cortes, V. (2003) "Lexical bundles in speech and writing: An initial taxonomy". In Wilson, A., Rayson, P. and McEnery, T. (eds.) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech.* Frankfurt am Main: Peter Lang, pp. 71-92.

Biber, D., Conrad, S. and Cortes, V. (2004) "*If you look at...*: Lexical bundles in university teaching and textbooks". In *Applied Linguistics* 25(3), 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Blake, N.F. (2002) *A Grammar of Shakespeare's Language*. Basingstoke and New York: Palgrave.

BNC = The British National Corpus. See Aston, G. and Burnard, L. (1998), Burnard L. (2000) and http://www.natcorp.ox.ac.uk/corpus/ (last accessed 11.08.12).

Bondi, M. (2010) "Perspectives on keywords and keyness". In Bondi, M. and Scott, M. (eds.), pp. 1-18.

Bondi, M. and Scott, M. (eds.) (2010) *Keyness in Texts*. Amsterdam and Philadelphia: John Benjamins.

Booth, S. (2004) "Shakespeare's language and the language of Shakespeare's time". In Alexander, C.M.S. (ed.), pp. 18-43. [First published in *Shakespeare Survey* 50, 1997].

Boyce, C. (1990) *Shakespeare A to Z. The Essential Reference to His Plays, His Poems, His Life and Times, and More.* New York: Dell Publishing.

Braunmuller, A.R. and Hattaway, M. (eds.) (2003) *The Cambridge Companion to English Renaissance Drama.* 2nd Edition. Cambridge: Cambridge University Press.

Bremond, C. (1966) "La logique des possible narratives". In *Communications* 4, 4-32.

Bremond, C. (1973) *Logique du Recit.* Paris: Seuil.

Burkert, M. (2011) "Tokens of Impersonation in Dekker's City Comedies". Web log post 19.11.11. See http://winedarksea.org/?paged=2 (last accessed 11.08.12).

Burnard, L. (2000) *Reference Guide for the British National Corpus: World Edition.* Oxford: Oxford University Computing Services. See also the 2007 online edition at http://www.natcorp.ox.ac.uk/docs/URG/ (last accessed 11.08.12).

Burrows, J.F. (1987) *Computation into Criticism: A Study of Jane Austen's Novels and Experiment in Method.* Oxford: Clarendon Press.

Burrows, J.F. (1992) "Computers and the study of literature". In Butler, C.S. (ed.) *Computers and Written Texts*. Oxford: B. Blackwell, pp. 167-204.

Burrows, J.F. (2007) "All the way through: Testing for authorship in different frequency strata". In *Literary and Linguistic Computing* 22(1), 27-47.

Busse, B. (2006) *Vocative Constructions in the Language of Shakespeare*. Amsterdam and Philadelphia: John Benjamins.

Busse, U. (2002a) *Linguistic Variation in the Shakespeare Corpus: Morpho-Syntactic Variability of Second Person Pronouns.* Amsterdam and Philadelphia: John Benjamins.

Busse, U. (2002b) "The co-occurrence of nominal and pronominal address forms in the Shakespeare Corpus: Who says *thou* or *you* to whom?" In Taavitsainen, I. and Jucker, A.H. (eds.), pp. 193-221.

Butler, M. (2003) "Private and occasional drama". In Braunmuller, A.R. and Hattaway, M. (eds.), pp. 131-63.

Calderwood, J.L. (1983) *TO BE AND NOT TO BE. Negation and Metadrama in* Hamlet. New York: Columbia University Press.

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

Cerasano, S.P. (2002) "Must the devil appear?: Audiences, actors, stage business". In Kinney, A.F. (ed.), pp. 193-211.

Chafe, W.L. (1982) "Integration and involvement in speaking, writing and oral literature". In Tannen, D. (ed.) *Spoken and Written Language: Exploring Orality and Literacy.* Norwood, New Jersey: Ablex, pp. 35-54.

Cockerham, H. (1623) *The English Dictionarie: Or, An Interpreter of hard English Words. Enabling as well Ladies and Gentlewomen, young Schollers, Clarkes, Merchants, as also Strangers of any Nation, to the vnderstanding of the more difficult Authors already printed in our Language, and the more speedy attaining of an elegant perfection of the English tongue, both in reading, speaking and writing. Being a Collection of the choisest words contained in the Table Alphabeticall and English Expositor, and of some thousands of words neuer published by any heretofore.* London: N. Butter.

Cook, A.J. (1997) "Audiences: investigation, interpretation, invention". In Cox, J.D. and Kastan, D.S. (eds.) *A New History of Early English Drama.* New York: Columbia University Press, pp. 305-20.

Craig, H. (1999) "Jonsonian chronology and the styles of *A Tale of a Tub*". In Butler, M. (ed.) *Presenting Ben Jonson: Text, History, Performance.* Basingstoke: Macmillan Press. and New York: St Martin's Press, pp. 210-32.

Craig, H. (2004) "Stylistic analysis and authorship studies". In Schreibman, S. Siemens, R. and Unsworth, J. (eds.) *A Companion to Digital Humanities.* Oxford: Blackwell, pp. 273-88.

Craig, H. (2010) "Style, statistics, and new models of authorship". In *Early Modern Literary Studies* 15(1), online edition. See http://purl.oclc.org/emls/15-1/craistyl.htm (last accessed 10.08.12).

Craig, H. (2011) "Shakespeare's vocabulary: Myth and reality". In *Shakespeare Quarterly* 62, 53-74.

Craig, H. (2012) "What are we to make of the fact that Shakespeare is typical, not exceptional, in some general stylistic tests?" Paper given at *Shakespeare Inside-out: Depth, Surface, Meaning* (the British Shakespeare Association 10th Anniversary Conference), Lancaster University, U.K., 26 February 2012.

Craig, H. and Kinney, A.F. (eds.) (2009) *Shakespeare, Computers, and the Mystery of Authorship.* Cambridge: Cambridge University Press.

Craig, H. and Kinney, A.F. (2009a) "Introduction". In Craig, H. and Kinney, A.F. (eds.), pp. 1-14.

Craig, H. and Kinney, A.F. (2009b) "Methods". In Craig, H. and Kinney, A.F. (eds.), pp. 15-39.

Craig, H. and Whipp, R. (2010) "Old spellings, new methods: automated procedures for indeterminate linguistic data". In *Literary and Linguistic Computing* 25(1), 37-52.

Crystal, D. (2008) *Think On My Words. Exploring Shakespeare's Language.* Cambridge: Cambridge University Press.

Crystal, D. (2012) *Spell it Out: The Singular Story of English Spelling.* London: Profile Books.

Crystal, D. and Crystal, B. (2002) *Shakespeare's Words. A Glossary and Language Companion.* London: Penguin. See also the searchable online glossary at www.shakespeareswords.com (last accessed 10.08.12).

Crystal, D. and Crystal, B. (2005) *The Shakespeare Miscellany.* London: Penguin.

Culpeper, J. (1996) "Towards an anatomy of impoliteness". In *Journal of Pragmatics* 25, 349-67.

Culpeper, J. (1997) "(Im)politeness in dramatic dialogue". In Culpeper, J., Short, M. and Verdonk, P. (eds.) *Exploring the Language of Drama: From Text to Context.* London and New York: Routledge, pp. 83-95.

Culpeper, J. (2001) *Language and Characterisation: People in Plays and Other Texts.* Harlow: Pearson Education.

Culpeper, J. (2002) "Computers, language and characterisation: An analysis of six characters in *Romeo and Juliet*". In Melander-Marttala, U., Ostman, C. and Kytö, M. (eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium,* Association Suedoise de Linguistique Appliquée (ASLA), 15. Universitetstryckeriet: Uppsala, pp.11-30.

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

Culpeper, J. (2011) "A new kind of dictionary for Shakespeare's plays: An immodest proposal". In Ravassat, M. and Culpeper, J. (eds.), pp. 58-83.

Culpeper, J. and Archer, D. (2008) "Requests and directness in Early Modern English trial proceedings and play-texts, 1640–1760". In Jucker, A.H. and Taavitsainen, I. (eds.) *Speech Acts in the History of English.* Amsterdam and Philadelphia: John Benjamins, pp. 45-84.

Culpeper, J. and Demmen, J. (2011) "Development of play-texts: From manuscript to print". In Pahta, P. and Jucker, A.H. (eds.) *Communicating Early English Manuscripts.* Cambridge: Cambridge University Press, pp. 162-77.

Culpeper, J. and Kytö, M. (2000) "Lexical bundles in Early Modern English dialogues. A window into the speech-related language of the past". In Fanego, T., Méndez-Naya, B. and Seoane, E. (eds.) *Sounds, Words, Texts and Change.* Amsterdam and Philadelphia: John Benjamins, pp. 45-63.

Culpeper, J. and Kytö, M. (2010) *Speech in Writing: Explorations in Early Modern English Dialogues.* Cambridge: Cambridge University Press.

Culpeper, J. and McIntyre, D. (2006) "Drama: Stylistic aspects". In Brown, K. (ed.) *Encyclopedia of Language and Linguistics.* Volume 3. 2nd Edition. Oxford: Elsevier, pp. 772-85.

Cusack, B. (2004) "Shakespeare and the tune of the time". In Alexander, C.M.S. (ed.), pp. 101-21. [First published in *Shakespeare Survey* 23, 1970].

Demmen, J.E.J. (2007) *Key lexical bundles in Shakespeare: An analysis of their functions and stylistic effects in comedy, tragedy and history plays.* Unpublished undergraduate dissertation, Lancaster University, U.K.

Demmen, J.E.J. (2009) *Charmed and chattering tongues: Investigating the functions and effects of key word clusters in the dialogue of Shakespeare's female characters.* Unpublished MA dissertation, Lancaster University, U.K.

Dillon, J. (2003) "Shakespeare and the traditions of English stage comedy". In Dutton, R. and Howard, J.E. (eds.), pp. 4-22.

Dunning, T. (1993) "Accurate methods for the statistics of surprise and coincidence". In *Computational Linguistics* 19(1), 61-74.

Dutton, R. (1991) *Mastering the Revels.* Basingstoke and London: Macmillan.

Dutton, R. (2000) *Licensing, Censorship and Authorship in Early Modern England.* Basingstoke and New York: Palgrave.

Dutton, R. (2011) "*Henry V*, January 7 1605". Paper given at *The Price of Peace* (in conjunction with the *Northern Renaissance Seminar*), Lancaster University, U.K., 10 June 2011.

Dutton, R. and Howard, J.E. (eds.) (2003) *A Companion to Shakespeare's Works.* Volume III. *The Comedies.* Malden, U.S.A, Oxford, U.K and Victoria, Australia: Blackwell.

EEBO = Early English Books Online, 1475-1700. ProQuest LLC. See http://eebo.chadwyck.com (last accessed 10.08.12).

Elliott, W.E.Y. and Valenza, R.J. (2011) "Shakespeare's vocabulary: Did it dwarf all others?" In Ravassat, M. and Culpeper, J. (eds.), pp. 34-57.

Enkvist, N.E. (1964) "On defining style". In Enkvist, N.E., Spencer, J. and Gregory, M. (eds.) *Linguistics and Style*. Oxford: Oxford University Press, pp. 1-56.

Erne, L. (2003) *Shakespeare as Literary Dramatist*. Cambridge: Cambridge University Press.

Evans, G. B. and Tobin, J.J.M. (1997) "Chronology and sources". In Evans, G.B. (ed., with Tobin, J.J.M.) *The Riverside Shakespeare*. 2nd Edition. Boston: Houghton-Mifflin, pp. 77-87.

Ferguson, C.A. (1981) "The structure and use of politeness formulas". In Coulmas, F. (ed.) *Conversational Routine. Explorations in Standardized Communication Situations and Prepatterned Speech.* The Hague: Mouton, pp. 21-35.

Findlay, A. (1999) *A Feminist Perspective on Renaissance Drama*. Oxford, U.K. and Malden, U.S.A.: Blackwell.

Fish, S. (1980) *Is There a Text in this Class? The Authority of Interpretive communities.* Cambridge, MA, U.S.A.: Harvard University Press.

Fish, S. E. (1996) "What is stylistics and why are they saying such terrible things about it?" In Weber, J.J. (ed.) *The Stylistics Reader.* London: Arnold, pp. 94-116.

Fish, S. (2012) "Mind your P's and B's: The digital humanities and interpretation". In *The New York Times*, 23 January 2012, The Opinion Pages. See http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/?scp=1&sq=stanlet%20fish&st=cse (accessed 21.3.12).

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

Fischer-Starcke, B. (2010) *Corpus Linguistics in Literary Analysis. Jane Austen and her Contemporaries.* London and New York: Continuum.

Fletcher, W.H. (2002-2007). *kfNgram*. KWiCFinder.com. See http://www.kwicfinder.com/kfNgram/kfNgramHelp.html (last accessed 10.08.12).

Foakes, R.A. (2003) "Playhouses and players". In Braunmuller, A.R. and Hattaway, M. (eds.), pp. 1-52.

Fraser, B. (1975) "Hedged performatives". In Cole, P. and Morgan, J.L. (eds.) *Syntax and Semantics.* Volume 3. *Speech Acts*. New York: Academic Press, pp. 187-210.

Freeman, D.C. (1995) "'Catch[ing] the nearest way': Macbeth and cognitive metaphor". In *Journal of Pragmatics* 24, 689-708.

Gabrielatos, C. and Baker, P. (2008) "Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005". In *Journal of English Linguistics* 36(1), 5-38.

Gabrielatos, C. and Marchi, A. (2011) "Keyness: Matching metrics to definitions". Paper given at *Corpus Linguistics in the South*, University of Portsmouth, U.K., 5 November 2011.

Galey, A. and Siemens, R. (2008) "Introduction: Reinventing Shakespeare in the digital humanities". In *Shakespeare* 4(3), 201-7.

Goffman, E. (1971) *Relations in Public. Microstudies of the Public Order.* London: Allen Lane.

Goodland, G. (2011) "'Strange deliveries': Contextualizing Shakespeare's first citations in the *OED*". In Ravassat, M. and Culpeper, J. (eds.), pp. 8-33.

Görlach, M. (1991) *Introduction to Early Modern English.* Cambridge: Cambridge University Press.

Greenblatt, S., Cohen, W., Howard, J.E. and Maus, K.E. (eds.) (1997) *The Norton Shakespeare.* London and New York: W.W. Norton and Company.

Greg, W.W. (ed.) (1908) *Henslowe's Diary.* 2 volumes. London: A.H. Bullen.

Grieve, J. (2007) "Quantitative authorship attribution: An evaluation of techniques". In *Literary and Linguistic Computing* 22(3), 251-70.

Gurr, A. (2000) "Maximal and minimal texts: Shakespeare v. the Globe". In *Globe Research Bulletin* 4, 1-25.

Halliday, M.A.K. (1978) *Language as Social Semiotic. The Social Interpretation of Language and Meaning.* London: Edward Arnold.

Halliday, M.A.K. (1994) *An Introduction to Functional Grammar.* 2nd Edition. London and New York: Edward Arnold.

Hammond, R. (ed.) (2009) *Double Falsehood* [Arden Shakespeare, 3rd Series Edition]. London: Methuen Drama, A & C Black Publishers.

Hapgood, R. (2004) "Shakespeare's thematic modes of speech: *Richard II* to *Henry VI*". In Alexander, C.M.S. (ed.), pp. 139-50. [First published in *Shakespeare Survey* 20, 1967].

Hardie, A. (in preparation) *Scripting for Linguists: An Introduction.* See http://www.lancs.ac.uk/staff/hardiea/php/PHPbook.pdf (last accessed 17.07.12).

Hattaway, M. (2003) "Drama and society". In Braunmuller, A.R. and Hattaway, M. (eds.), pp. 93-130.

Heinemann, M. (2003) "Political drama". In Braunmuller, A.R. and Hattaway, M. (eds.), pp. 164-96.

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 Nevanlina and Irma Taavittsainen (Middle English); Terttu Nevalainen and Helena Raumolin-Brunberg (Early Modern English). Helsinki: Department of English, University of Helsinki.

Herman, V. (1995) *Dramatic Discourse. Dialogue as Interaction in Plays*. London and New York: Routledge.

Hidalgo Downing, L. (2000) *Negation, Text Worlds and Discourse: The Pragmatics of Fiction.* Volume LXVI. *Advances in Discourse Processes*. Stamford, CT, U.S.A: Ablex.

Hillman, R. (1997) *Self-Speaking in Medieval and Early Modern English Drama. Subjectivity, Discourse and the Stage*. Basingstoke: Macmillan Press and New York: St. Martin's Press.

Hilpert, M. and Gries, S.T. (2009) "Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition". In *Literary and Linguistic Computing* 24(4), 385-401.

Ho, Y. (2011) *Corpus Stylistics in Principles and Practice. A Stylistic Exploration of John Fowles' The Magus*. London and New York: Continuum.

Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hoey, M. (2007) "Lexical priming and literary creativity". In Hoey, M., Mahlberg, M., Stubbs, M. and Teubert, W., pp. 7-29.

Hoey, M., Mahlberg, M., Stubbs, M. and Teubert, W. (2007) *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum.

Hoover, D.L. (2010) "Authorial style". In McIntyre, D. and Busse, B. (eds.), pp. 250-92.

Hope, J. (1994) *The Authorship of Shakespeare's Plays*. Cambridge: Cambridge University Press.

Hope, J. (2004) "Shakespeare and language: An introduction." In Alexander, C.M.S. (ed.), pp. 1-17.

Hope, J. (2010) *Shakespeare and Language: Reason, Eloquence and Artifice in the Renaissance*. London: Arden Shakespeare.

Hope, J. (2011) "The digital renaissance: Mapping the language of drama 1550-1700". Presentation given at the symposium *'Set the Word Itself Against the Word': New Directions in Early Modern Textual Analysis,* Lancaster University, U.K., 15 October 2011.

Hope, J. and Witmore, M. (2004) "The very large textual object: A prosthetic reading of Shakespeare". In *Early Modern Literary Studies* 9(3) Special Issue 12, 6.1-36. See http://extra.shu.ac.uk/emls/09-3/hopewhit.htm (accessed 17.02.12).

Hope, J. and Witmore, M. (2010) "The hundredth psalm to the tune of 'Green Sleeves': Digital approaches to the language of genre". In *Shakespeare Quarterly* 61(3), 357-90.

Hopkins, L. (2004) "Household words: Macbeth and the failure of spectacle". In Alexander, C.M.S. (ed.), pp. 251-65. [First published in *Shakespeare Survey* 50, 1997].

Hori, M. (2004) *Investigating Dickens' Style. A Collocational Analysis*. Basingstoke and New York: Palgrave Macmillan.

Horn, L.R. (2001) *A Natural History of Negation.* Stanford, CA, U.S.A.: CSLI Publications.

Hunston, S. (2002) *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hunter, G.K. (1997) *English Drama 1586-1642. The Age of Shakespeare* [The Oxford History of English Literature]. Oxford: Oxford University Press.

Ihaka, R. and Gentleman, R. (1996) "R: A language for data analysis and graphics". In *Journal of Computational and Graphical Statistics* 5, 299-314.

Inaki, A. and Okita, T. (2006) "A small-corpus-based approach to Alice's roles". In *Literary and Linguistic Computing* 21(3), 283-94.

Ingram, R. and Ingram, M. (2012) "Diachronic variables, proximity, and distance in the verse dramas of Shakespeare and his contemporaries". Paper given at *Shakespeare Inside-out: Depth, Surface, Meaning* (the British Shakespeare Association 10th Anniversary Conference), Lancaster University, U.K., 26 February 2012.

Ioppolo, G. (2002) "The transmission of a play-text". In Kinney, A.F. (ed.), pp. 163-79.

Jakobson, R. (1960) "Closing statement: linguistics and poetics". In Sebeok, T.A. (ed.) *Style in Language.* Cambridge, MA, U.S.A.: M.I.T. Press, pp. 350-57.

Jardine, L. (1983) *Still Harping on Daughters: Women and Drama in the Age of Shakespeare.* Sussex, U.K.: Harvester Press and New Jersey, U.S.A.: Barnes and Noble.

Jeffries, L. and McIntyre, D. (2010) *Stylistics.* Cambridge: Cambridge University Press.

Jeffries, L. and Walker, B. (2012) "Key words in the press". In *English Text Construction* 5(2), 208-29.

Jespersen, O. (1917) "Negation in English and other languages". In *Historisk-filologiske Meddeleser* 1, pp. 1-151. Reprinted in *Selected Writings of Otto Jespersen*, London and Tokyo: George Allen & Unwin and Sejo Publishing Co., 1962.

Jockers, M.L. and Witten, D.M. (2010) "A comparative study of machine learning methods for authorship attribution". In *Literary and Linguistic Computing* 25(2), 215-23.

Jucker, A. H. (2002) "Discourse markers in Early Modern English". In Watts, R. and Trudgill, P. (eds.) *Alternative Histories of English.* London: Routledge, pp. 210-30.

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. See http://corp.hum.sdu.dk/cqp.en.html (last accessed 19.03.12).

Kermode, F. (2000) *Shakespeare's Language.* London: Penguin.

Kinney, A.F. (ed.) (1999) *Renaissance Drama. An Anthology of Plays and Entertainments.* Malden, U.S.A., Oxford, U.K., Melbourne, Australia and Berlin, Germany: Blackwell.

Kinney, A.F. (ed.) (2002) *A Companion to Renaissance Drama.* Oxford, U.K. and Malden, U.S.A.: Blackwell.

Kinney, A.F. (2009) "Authoring *Arden of Faversham*". In Craig, H. and Kinney, A.F. (eds.), pp. 78-115.

Knutson, R.L. (2002) "Playing companies and repertory". In Kinney, A.F. (ed.), pp. 180-92.

Koller, V., Hardie, A., Rayson, P. and Semino, E. (2008). "Using a semantic annotation tool for the analysis of metaphor in discourse". In *metaphorik.de* 15/2008. See http://www.metaphorik.de/15/koller.pdf (last accessed 10.08.12).

Kryk-Kastovsky, B. (2000) "Representations of orality in Early Modern English trial records". In *Journal of Historical Pragmatics* 1(2), 201-30.

Kytö, M. (comp.)(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, M. and Walker, T. (2006) *Guide to A Corpus of English Dialogues 1560-1760.* Acta Universitatis Upsaliensis. Studia Anglistica Upsaliensia 130. Uppsala: Uppsala Universitet.

LION = Literature Online (1996-2012). Proquest LLC. See http://lion.chadwyck.co.uk/searchFullrec.do?id=R03650648&area=mla&forward=critref_fr&DurUrl=Yes (accessed 31.08.12).

Lakoff, G. and Johnson, M. (1980) *Metaphors We Live By.* Chicago and London: University of Chicago Press.

Lambrou, M. and Stockwell, P. (eds.) (2007) *Contemporary Stylistics.* London and New York: Continuum.

Lambrou, M. and Stockwell, P. (2007) "Introduction: The state of contemporary stylistics". In Lambrou, M. and Stockwell, P. (eds.), pp. 1-4.

Lancashire, I. (ed.) (2012) *Lexicon of Early Modern English.* University of Toronto Press. Online resource. See http://leme.library.utoronto.ca/public/intro.cfm (accessed 23.03.12).

Leech, G. (1985) "Stylistics". In van Dijk, T.A. (ed.) *Discourse and literature.* Amsterdam and Philadelphia: John Benjamins, pp. 39-57.

Leech, G. (1991) "The state of the art in corpus linguistics". In Aimer, K. and Altenberg, B. (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* London: Longman, pp. 105-22.

Leech, G. (2008) *Language in Literature. Style and Foregrounding.* Harlow: Pearson Education Limited.

Leech, G. (2012) "Getting up to date with the Brown family of corpora: BE06 and AmE06, and what they can tell us about contemporary grammatical change". Paper given at the *UCREL Corpus Research Seminar*, Lancaster University, U.K., 15 March 2012.

Leech, G.N., Garside, R. & Bryant, M. (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. See http://ucrel.lancs.ac.uk/claws/ (last accessed 10.08.12).

Leech, G., Hundt, M., Mair, C. and Smith, N. (2009) *Change in Contemporary English: A Grammatical Study.* Cambridge: Cambridge University Press.

Leech, G., Rayson, P. and Wilson, A. (2001) *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* London: Longman.

Leech, G. and Short, M. (2007) *Style in Fiction: A Linguistic Introduction to English Fictional Prose.* 2nd Edition. Harlow: Pearson Education.

Leech, G. and Smith, N. (2005) "Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB". In *ICAME Journal* 29, 83-98.

Leggatt, A. (1988) *English Drama: Shakespeare to the Restoration, 1590-1660* [Longman Literature in English Series]. Harlow and New York: Longman.

Leggatt, A. (1999) *Introduction to English Renaissance Comedy.* Manchester and New York: Manchester University Press.

Lehto, A., Baron, A., Ratia, M. and Rayson, P. (2010) "Improving the precision of corpus methods: The standardized version of Early Modern English Medical Texts". In Taavitsainen, I. and Pahta, P. (eds.) *Early Modern English Medical Texts: Corpus Description and Studies.* Amsterdam: John Benjamins, pp. 279-90.

Levenson, J. (2003) "Comedy". In Braunmuller, A.R. and Hattaway, M. (eds.) pp. 254-91.

Levinson, S.C. (1983) *Pragmatics.* Cambridge: Cambridge University Press.

Louw, B. (2008) "Consolidating empirical method in data-assisted stylistics. Towards a corpus-attested glossary of literary terms". In Zyngier, S., Bortolussi, M., Chesnokova, A. and Auracher, J. (eds.) *Directions in Empirical Literary Studies.* Amsterdam and Philadelphia: John Benjamins, pp. 243–64.

Lutzky, U. (2009a) *Discourse markers in Early Modern English: The case of 'marry', 'well' and 'why'.* Unpublished PhD thesis, University of Vienna, Austria.

Lutzky, U. (2009b) *The 'Drama Subcorpus' – Manual.* Unpublished document provided by the author.

Lutzky, U. (2012) *Discourse Markers in Early Modern English.* Amsterdam: John Benjamins.

Lutzky, U. and Demmen, J. (forthcoming) "*Pray* in Early Modern English drama". In *Journal of Historical Pragmatics* 14(2).

Mahlberg, M. (2007) "Clusters, key clusters and local textual functions in Dickens". In *Corpora* 2(1), 1-31.

Mahlberg, M. (2009) "Corpus stylistics and the Pickwickian watering-pot". In Baker, P. (ed.) *Contemporary Corpus Linguistics.* London: Continuum, pp. 47-63.

Mahlberg, M. and McIntyre, D. (2011) "A case for corpus stylistics: Ian Fleming's Casino Royale". In *English Text Construction* 4(2), 204-27.

Mahlberg, M. and Smith, C. (2010) "Corpus approaches to prose fiction: Civility and body language in *Pride and Prejudice*". In McIntyre, D. and Busse, B. (eds.), pp. 449-67.

Marche, S. (2011) *How Shakespeare Changed Everything*. New York: HarperCollins.

Mazzon, G. (2002) "Pronouns and nominal address in Shakespearean English. A socio-affective marking system in transition". In Taavitsainen, I. and Jucker, A.H. (eds.), pp. 223-307.

Mazzon, G. (2009) *Interactive Dialogue Sequences in Middle English Drama.* Amsterdam and Philadelphia: John Benjamins.

McEnery, T. (2009) "Keywords and moral panics: Mary Whitehouse and media censorship". In Archer, D. (ed.), pp. 93-124.

McEnery, T. and Hardie, A. (2011) *Corpus Linguistics: Method, Theory and Practice.* Cambridge and New York: Cambridge University Press.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics: An Introduction.* 2nd Edition. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies: An Advanced Resource Book.* Abingdon and New York: Routledge.

McIntyre, D. (2010) "Dialogue and characterization in Quentin Tarantino's *Reservoir Dogs*: A corpus stylistic analysis". In McIntyre, D. and Busse, B. (eds.), pp. 162-83.

McIntyre, D. and Busse, B. (eds.) (2010) *Language and Style*. Basingtoke and New York: Palgrave Macmillan.

McIntyre, D. and Walker, B. (2011) "Discourse presentation in Early Modern English writing. A preliminary corpus-based investigation". In *International Journal of Corpus Linguistics* 16(1), 101-30.

McRae, A. (2003) *Renaissance Drama*. London: Arnold.

Mitchell, E.R. (1971) *Pronouns of address in English, 1580-1780: A study of form changes as reflected in British drama.* Unpublished PhD thesis, Texas A & M University, U.S.A.

Moon, R. (1998) *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford: Clarendon Press.

Morley, J. and Partington, A. (2009) "A few *frequently asked questions* about semantic – or *evaluative* – prosody*". In *International Journal of Corpus Linguistics* 14(2), 139-58.

Mueller, M. (2005) "The Nameless Shakespeare". In *TEXT Technology* (No. 1, 2005): 61-70. See http://texttechnology.mcmaster.ca/pdf/vol14_1_06.pdf (last accessed 10.08.12).

Mueller, M., Parod, W., Cousens, J., Burns, P. and Norstad, J. (2006). *WordHoard.* Evanston, Illinois: Northwestern University. See http://wordhoard.northwestern.edu/userman/whatiswordhoard.html (last accessed 10.08.12).

Mukařovský, J. (1964a) "Standard language and poetic language". In Garvin, P.L. (ed.) *A Prague School Reader on Aesthetics, Literary Structure and Style.* Washington: Georgetown University Press, pp. 17-30.

Mukařovský, J. (1964b) "The esthetics of language". In Garvin, P.L. (ed.) *A Prague School Reader on Aesthetics, Literary Structure and Style.* Washington: Georgetown University Press, pp. 31-69.

Munro, L. (2005) *Children of the Queen's Revels: A Jacobean Theatre Repertory.* Cambridge: Cambridge University Press.

Murphy, S. (2007) "Now I am alone: A corpus stylistic approach to Shakespearian soliloquies". In Gabrielatos, C., Slessor, R. and Unger, J. (eds.) *Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching,* Volume 1. *Papers from LAEL PG 2006,* online edition. See http://www.ling.lancs.ac.uk/pgconference/v01/Volume01.pdf (last accessed 11.08.12).

Nevalainen, T. (2006) *An Introduction to Early Modern English.* Edinburgh: Edinburgh University Press.

Nevalainen, T. and Raumolin-Brunberg, H. (2003) *Historical Sociolinguistics.* Harlow: Pearson Education.

Notepad++ (V.5.6.8, 2010). GNU General Public License. See http://notepad-plus-plus.org/ (last accessed 31.08.12).

O'Halloran, K. (2007) "Corpus-assisted literary evaluation". In *Corpora* 2(1), 33-63.

Oakes, M.P. (1998) *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

OED = The Oxford English Dictionary (1989) 2nd Edition. Oxford: Oxford University Press. Online version http://www.oed.com/ (last accessed 10.08.12).

Oncins-Martínez, J.L. (2006) "Notes on the metaphorical basis of sexual language in Early Modern English". In Vázquez-González, J.L., Martínez Vázquez, M. and Ron Vaz, P. (eds.) *The Historical Linguistics-Cognitive Linguistics Interface.* Huelva, Spain: University of Huelva Press, pp. 205-24.

Oncins-Martínez, J.L. (2011) "Shakespeare's sexual language and metaphor: A cognitive-stylistic approach". In Ravassat, M. and Culpeper, J. (eds.), pp. 215-45.

Onions, C.T. (1982) *A Shakespeare Glossary*. 2nd Edition. Oxford: Clarendon Press.

Orgel, S. (1996) *Impersonations. The Performance of Gender in Shakespeare's England.* Cambridge: Cambridge University Press.

Orlin, L.C. (2003) "Shakespearean comedy and material life". In Dutton, R. and Howard, J.E. (eds.), pp. 159-81.

PHP (2001-2012). The PHP Group. See http://www.php.net/ (last accessed 31.08.12).

Pennebaker, J.W. (2011) *The Secret Life of Pronouns: What Our Words Say About Us.* New York: Bloomsbury Press.

Petersen, L.B. (2010) *Shakespeare's Errant Texts. Textual Form and Linguistic Style in Shakespearean 'Bad' Quartos and Co-authored Plays.* Cambridge: Cambridge University Press.

Piao, W., Wilson, A., and McEnery, A. (2002) "A multilingual corpus toolkit". Paper given at the *Fourth North American Symposium on Corpus Linguistics*, 1-3 November 2002, Indianapolis, Indiana. See also https://sites.google.com/site/scottpiaosite/software/mlct (last accessed 10.08.12).

Project Gutenberg. Online text archive. Project Gutenberg & PROMO.NET. See http://www.gutenberg.org/ebooks/ (last accessed 10.08.12).

Ravassat, M. and Culpeper, J. (eds.) (2011) *Stylistics and Shakespeare's Language. Transdisciplinary Approaches.* London and New York: Continuum.

Rackin, P. (2003) "Shakespeare's crossdressing comedies". In Dutton, R. and Howard, J.E. (eds.), pp. 114-36.

Rayson, P. (2008) "From key words to key semantic domains". In *International Journal of Corpus Linguistics* 13(4), 519-549.

Rayson, P. (2009) *Wmatrix: A web-based corpus processing environment.* Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/ (last accessed 10.08.12).

Rayson, P., Archer, D., Piao, S.L. & McEnery, T. (2004a) "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. Paris: European Language Resources Association, pp. 7-12.

Rayson, P., Archer, D., Smith, N. (2005) "VARD versus WORD: A comparison of the UCREL variant detector and modern spellcheckers on English historical corpora". In *Proceedings of Corpus Linguistics 2005*, Birmingham University, U.K., 14-17 July 2005. See http://ucrel.lancs.ac.uk/VariantSpelling/ (last accessed 10.08.12).

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007) "Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora". In *Proceedings of Corpus Linguistics 2007*, Birmingham University, U.K., 27-30 July 2007. See http://ucrel.lancs.ac.uk/VariantSpelling/ (last accessed 10.08.12).

Rayson, P., Berridge, D. and Francis, B. (2004b) "Extending the Cochran rule for the comparison of word frequencies between corpora". In Purnelle, G., Fairon, C. and Dister, A. (eds.) *Le poids des mots: Proceedings of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*. Volume II. pp. 926-936. 10-12 March 2004. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain. See http://www.comp.lancs.ac.uk/~paul/publications/rbf04_jadt.pdf (last accessed 11.08.12).

Replogle, C.A. (1967) *Shakespeare's use of the forms of address*. Unpublished dissertation, Brandeis University, U.S.A.

Richmond, H.M. (2002) *Shakespeare's Theatre: A Dictionary of his Stage Context.* London and New York: Continuum. Online edition. See http://www.credoreference.com (last accessed 10.08.12).

Rosso, O.A., Craig, H. and Moscato, P. (2009) "Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers". In *Physica A* 388(6), 916-26.

Salmon, V. (1999) "Orthography and punctuation". In Lass, R. (ed.) *The Cambridge History of the English Language.* Volume III. *1476-1776*. Cambridge: Cambridge University Press, pp. 13-55.

Schmitt, N. and Carter, R. (2004) "Formulaic sequences in action: An introduction". In Schmitt, N. (ed.) *Formulaic Sequences.* Amsterdam and Philadelphia: John Benjamins, pp. 1-22.

Schneider, P. (2002) "Computer assisted spelling normalization of 18th century English". In Peters, P, Collins, P. and Smith, A. (eds.) *New Frontiers of Corpus Research: Papers from the Twenty-First International Conference on English Language Research on Computerized Corpora.* Amsterdam: Rodopi, pp. 199-211.

Scott, M. (1997) "PC analysis of key words – and key key words". In *System* 25(2), 233-45.

Scott, M. (1999) *WordSmith Tools Version 3.0*. Oxford: Oxford University Press. See http://www.lexically.net/wordsmith/index.html (last accessed 08.08.12).

Scott, M. (1999:Help menu) *WordSmith Tools Help*. Liverpool: Lexical Analysis Software.

Scott, M. (2000) "Focusing on the text and its key words". In Burnard, L. and McEnery, T. (eds.) *Rethinking Language Pedagogy from a Corpus Perspective. Papers from the Third International Conference on Teaching and Language Corpora*, Volume 2. Frankfurt am Main: Peter Lang, pp. 103-21.

Scott, M. (2009) "In search of a bad reference corpus". In Archer, D. (ed.), pp. 79-91.

Scott, M. (2010) "Problems in investigating keyness". In Bondi, M. and Scott, M. (eds.), pp. 43-57.

Scott, M. and Thompson, G. (eds.) (2001) *Patterns of Text in Honour of Michael Hoey*. Amsterdam and Philadelphia: John Benjamins.

Scott, M. and Tribble, C. (2006) *Textual Patterns. Key Words and Corpus Analysis in Language Education.* Amsterdam and Philadelphia: John Benjamins.

Scragg, D.G. (2011 [1974]) *A History of English Spelling*. New Edition [Mont Follick Series]. Manchester: Manchester University Press.

Searle, J. R. (1969) *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Semino, E. and Culpeper, J. (2011) "Stylistics". In Östman, J.-A. and Verschueren, J. (eds.) *Pragmatics in Practice* [Handbooks of Pragmatics Highlights 9]. Amsterdam and Philadelphia: John Benjamins, pp. 295-305.

Semino, E. and Short, M. (2004) *Corpus Stylistics*. Abingdon, U.K. and New York, U.S.A.: Routledge.

Shapiro, M. (2002) "Boy companies and private theatres". In Kinney, A.F. (ed.), pp. 314-25.

Short, M. (1996) *Exploring the Language of Poems, Plays and Prose*. London and New York: Longman.

Sinclair, J. (1966) "Beginning the study of lexis". In Bazell, C., Halliday, M.A.K., Robins, R.H. and Catford, J. (eds.) *In Memory of J.R. Firth.* London: Longman, pp. 410-30.

Sinclair, J. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J.M. (edited with Carter, R.) (2004) *Trust the Text: Language, Corpus and Discourse*. London and New York: Taylor and Francis.

Spevack, M. (1968-80) *A Complete and Systematic Concordance to the Works of Shakespeare.* 9 volumes. Hildesheim: George Ohms.

SPSS. Predictive Analytics Software. IBM. See http://www-01.ibm.com/software/uk/analytics/spss/ (accessed 12.09.12).

Stein, D. (2002) "Pronomial usage in Shakespeare: Between sociolinguistics and conversation analysis". In Taavitsainen, I. and Jucker, A.H. (eds.), pp. 251-307.

Stern, T. (2000) *Rehearsal from Shakespeare to Sheridan.* Oxford and New York: Oxford University Press.

Stern, T. (2004) "Re-patching the play". In Holland, P. and Orgel, S. (eds.) *From Script to Stage in Early Modern England.* Basingstoke: Palgrave Macmillan, pp. 151-80.

Stubbs, M. (1996) *Text and Corpus Analysis.* Oxford, U.K. and Malden, U.S.A.: Blackwell.

Stubbs, M. (2001) *Words and Phrases. Corpus Studies of Lexical Semantics.* Oxford, U.K. and Malden, U.S.A.: Blackwell.

Stubbs, M. (2005) "Conrad in the computer: Examples of quantitative stylistic methods". In *Language and Literature* 14(1), 5-24.

Stubbs, M. (2007) "Quantitative data on multi-word sequences in English: The case of the word *world*". In Hoey, M., Mahlberg, M., Stubbs, M. and Teubert, W., pp. 163-89.

Stubbs, M. (2010) "Three concepts of keywords". In Bondi, M. and Scott, M. (eds.), pp. 21-42.

Stubbs, M. and Barth, I. (2003) "Using recurrent phrases as text-type discriminators. A quantitative method and some findings". In *Functions of Language* 10(1), 61–104.

Sullivan, G.A., Jr. (2006) "Shakespeare's comic geographies". In Dutton, R. and Howard, J.E. (eds.), pp. 182-99.

Taavitsainen, I. (1999) "Personality and styles of affect in *The Canterbury Tales*". In Lester, G. (ed.) *Chaucer in Perspective: Middle English Essays in Honour of Norman Blake.* Sheffield: Sheffield Academic Press, pp. 218-34.

Taavitsainen, I. and Jucker, A.H. (eds.) (2002) *Diachronic Perspectives on Address Term Systems.* Amsterdam and Philadelphia: John Benjamins.

Talbot, M.M. (2010) *Language and Gender.* 2nd Edition. Cambridge: Polity Press.

Tieken-Boone van Ostade, I., Tottie, G. and van der Wurff, W. (eds.) (1998) *Negation in the History of English.* Berlin and New York: Mouton de Gruyter.

Tissari, H. (2009) "Soul-searching in Shakespeare". In Tissari, H. (ed.) *Approaches to Language and Cognition* [Studies in Variation, Contacts and Change in English. Volume 3]. See http://www.helsinki.fi/varieng/journal/volumes/03/tissari/ (accessed 19.01.12).

Tissari, H. (2010a) "Love, metaphor and responsibility: Some examples from Early Modern and Present-Day English corpora." In Low, G., Todd, Z., Deignan, A. and Cameron, L. (eds.) *Researching and Applying Metaphor in the Real World* [Human Cognitive Processing 26]. Amsterdam: John Benjamins, pp. 125-43.

Tissari, H. (2010b) "English words for emotions and their metaphors." In Winters, E., Tissari, H. and Allan, K. (eds.) *Historical Cognitive Linguistics*, pp. 298-329.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work.* Amsterdam and Philadelphia: John Benjamins.

Tottie, G. (1991) *Negation in English Speech and Writing. A Study in Variation* [Quantitative Analyses of Linguistic Structure 4]. San Diego and London: Academic Press.

Traugott, E. C., and Dasher, R.B. (2002) *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Tricomi, A.H. (2004) "The aesthetics of mutilation in *Titus Andronicus*". In Alexander, C.M.S. (ed.), pp. 226-39. [First published in *Shakespeare Survey* 27, 1974].

Tucker Brooke, C.F. (1908) *The Shakespeare Apocrypha, Being a Collection of Fourteen Plays Which Have Been Ascribed to Shakespeare.* Oxford: Clarendon Press.

Twyning, J.A. (2002) "City comedy". In Kinney, A.F. (ed.), pp. 353-66.

van der Wouden, T. (1997) *Negative Contexts. Collocation, Polarity and Multiple Negation* [Routledge Studies in Germanic Linguistics]. London and New York: Routledge.

van Peer, W. (1986) *Stylistics and Psychology: Investigations of Foregrounding.* London: Croom Helm.

van Peer, W. (1989) "Quantitative studies of literature: A critique and an outlook". In *Computers and the Humanities* 23(4/5), 301-7.

Vickers, B. (2002) *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays.* Oxford: Oxford University Press.

Wales, K. (2001) *A Dictionary of Stylistics*. 2nd Edition. Harlow: Pearson Education.

Walker, B. (2010) "Wmatrix, key concepts and the narrators in Julian Barnes's *Talking It Over*". In McIntyre, D. and Busse, B. (eds.), pp. 364-87.

Walker, B. (2012) *Character and characterisation in Julian Barnes'* Talking It Over: *A corpus stylistic analysis*. Unpublished PhD thesis, Lancaster University, U.K.

Walker, T. (2007) Thou *and* You *in Early Modern English Dialogues*. Amsterdam: John Benjamins.

Wall, W. (2003) "*The Merry Wives of Windsor*: Unhusbanding desires in Windsor". In Dutton, R. and Howard, J. E. (eds.), pp. 376-92.

Watson, R.N. (2003) "Tragedy". In Braunmuller, A.R. and Hattaway, M. (eds.), pp. 292-343.

Watt, T.I. (2009) "The authorship of *The Raigne of Edward the Third*". In Craig, H. and Kinney, A.F. (eds.), pp. 116-33.

Wells, S., Taylor, G. with Jowett, J. and Montgomery, W. (1987) *William Shakespeare: A Textual Companion.* Oxford: Clarendon Press.

Westfall, S. (2002) "'What revels are in hand?': Performances in the great households". In Kinney, A.F. (ed.), pp. 266-80.

Williams, R. (1976) *Keywords*. London: Fontana.

Wilson, A. (2011) "Embracing Bayes Factors for key item analysis in corpus linguistics". Paper given at *UCREL Summer School in Corpus Linguistics*, Lancaster University, U.K., 13-15 July 2011.

Witmore, M. (2012) "What did Stanley Fish count, and when did he start counting it?" Web log post 27.01.12. See http://winedarksea.org/ (accessed 21.03.12).

Włodarczyk, M. (2007) *Pragmatic Aspects of Reported Speech. The Case of Early Modern English Courtroom Discourse.* Frankfurt am Main: Peter Lang.

Wray, A. (2002) *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Wray, A. (2008) *Formulaic Language. Pushing the Boundaries.* Oxford: Oxford University Press.

Wray, A. (2009) "Identifying formulaic language. Persistent challenges and new opportunities". In Corrigan, R., Moravcsik, E.A., Ouali, H. and Wheatley, K.M. (eds.) *Formulaic Language.* Volume 1. *Distribution and Historical Change.* Amsterdam and Philadelphia: John Benjamins, pp. 27-51.

Wynne, M. (2006) "Stylistics: Corpus approaches". In Brown, K. (ed.) *Encyclopedia of Language and Linguistics.* Volume 12. 2nd Edition. Oxford: Elsevier, pp. 223-6.

Wynne, M. (2010) "Interdisciplinary relationships". In *International Journal of Corpus Linguistics* 15(3), 425-7.

Appendix I
USAS semantic tagset (all categories)

A GENERAL & ABSTRACT TERMS
A1 General
A1.1.1 General actions, making etc.
A1.1.2 Damaging and destroying
A1.2 Suitability
A1.3 Caution
A1.4 Chance, luck
A1.5 Use
A1.5.1 Using
A1.5.2 Usefulness
A1.6 Physical/mental
A1.7 Constraint
A1.8 Inclusion/Exclusion
A1.9 Avoiding
A2 Affect
A2.1 Affect: Modify, change
A2.2 Affect: Cause/Connected
A3 Being
A4 Classification
A4.1 Generally kinds, groups, examples
A4.2 Particular/general; detail
A5 Evaluation
A5.1 Evaluation: Good/bad
A5.2 Evaluation: True/false
A5.3 Evaluation: Accuracy
A5.4 Evaluation: Authenticity
A6 Comparing
A6.1 Comparing: Similar/different
A6.2 Comparing: Usual/unusual
A6.3 Comparing: Variety
A7 Definite (+ modals)
A8 Seem
A9 Getting and giving; possession
A10 Open/closed; Hiding/Hidden; Finding; Showing
A11 Importance
A11.1 Importance: Important
A11.2 Importance: Noticeability
A12 Easy/difficult
A13 Degree
A13.1 Degree: Non-specific
A13.2 Degree: Maximizers
A13.3 Degree: Boosters
A13.4 Degree: Approximators
A13.5 Degree: Compromisers
A13.6 Degree: Diminishers
A13.7 Degree: Minimizers
A14 Exclusivizers/particularizers
A15 Safety/Danger

B THE BODY & THE INDIVIDUAL
B1 Anatomy and physiology
B2 Health and disease
B3 Medicines and medical treatment
B4 Cleaning and personal care
B5 Clothes and personal belongings

C ARTS & CRAFTS
C1 Arts and crafts

E EMOTIONAL ACTIONS, STATES & PROCESSES
E1 General
E2 Liking
E3 Calm/Violent/Angry
E4 Happy/sad
E4.1 Happy/sad: Happy
E4.2 Happy/sad: Contentment
E5 Fear/bravery/shock
E6 Worry, concern, confident

F FOOD & FARMING
F1 Food
F2 Drinks
F3 Cigarettes and drugs
F4 Farming & Horticulture

G GOVT. & THE PUBLIC DOMAIN
G1 Government, Politics & elections
G1.1 Government etc.
G1.2 Politics
G2 Crime, law and order
G2.1 Crime, law and order: Law & order
G2.2 General ethics
G3 Warfare, defence and the army; Weapons

H ARCHITECTURE, BUILDINGS, HOUSES & THE HOME
H1 Architecture, kinds of houses & buildings
H2 Parts of buildings
H3 Areas around or near houses
H4 Residence
H5 Furniture and household fittings

I MONEY & COMMERCE
I1 Money generally
I1.1 Money: Affluence
I1.2 Money: Debts
I1.3 Money: Price
I2 Business
I2.1 Business: Generally
I2.2 Business: Selling
I3 Work and employment
I3.1 Work and employment: Generally
I3.2 Work and employment: Professionalism
I4 Industry

K ENTERTAINMENT, SPORTS & GAMES
K1 Entertainment generally
K2 Music and related activities
K3 Recorded sound etc.
K4 Drama, the theatre & show business
K5 Sports and games generally
K5.1 Sports
K5.2 Games
K6 Children's games and toys

L LIFE & LIVING THINGS
L1 Life and living things
L2 Living creatures generally
L3 Plants

M MOVEMENT, LOCATION, TRAVEL & TRANSPORT
M1 Moving, coming and going
M2 Putting, taking, pulling, pushing, transporting &c.
M3 Movement/transportation: land
M4 Movement/transportation: water
M5 Movement/transportation: air
M6 Location and direction
M7 Places
M8 Remaining/stationary

N NUMBERS & MEASUREMENT
N1 Numbers
N2 Mathematics
N3 Measurement
N3.1 Measurement: General
N3.2 Measurement: Size
N3.3 Measurement: Distance
N3.4 Measurement: Volume
N3.5 Measurement: Weight
N3.6 Measurement: Area
N3.7 Measurement: Length & height
N3.8 Measurement: Speed
N4 Linear order
N5 Quantities
N5.1 Entirety; maximum
N5.2 Exceeding; waste
N6 Frequency etc.

O SUBSTANCES, MATERIALS, OBJECTS & EQUIPMENT
O1 Substances and materials generally
O1.1 Substances and materials generally: Solid
O1.2 Substances and materials generally: Liquid
O1.3 Substances and materials generally: Gas
O2 Objects generally
O3 Electricity and electrical equipment
O4 Physical attributes
O4.1 General appearance and physical properties
O4.2 Judgement of appearance (pretty etc.)
O4.3 Colour and colour patterns
O4.4 Shape
O4.5 Texture
O4.6 Temperature

P EDUCATION
P1 Education in general

Q LINGUISTIC ACTIONS, STATES & PROCESSES
Q1 Communication
Q1.1 Communication in general
Q1.2 Paper documents and writing
Q1.3 Telecommunications
Q2 Speech acts
Q2.1 Speech etc: Communicative
Q2.2 Speech acts
Q3 Language, speech and grammar
Q4 The Media
Q4.1 The Media: Books
Q4.2 The Media: Newspapers etc.
Q4.3 The Media: TV, Radio & Cinema

S SOCIAL ACTIONS, STATES & PROCESSES
S1 Social actions, states & processes
S1.1 Social actions, states & processes
S1.1.1 General
S1.1.2 Reciprocity
S1.1.3 Participation
S1.1.4 Deserve etc.
S1.2 Personality traits
S1.2.1 Approachability and Friendliness
S1.2.2 Avarice
S1.2.3 Egoism
S1.2.4 Politeness
S1.2.5 Toughness; strong/weak
S1.2.6 Sensible
S2 People
S2.1 People: Female
S2.2 People: Male
S3 Relationship
S3.1 Relationship: General
S3.2 Relationship: Intimate/sexual
S4 Kin
S5 Groups and affiliation
S6 Obligation and necessity
S7 Power relationship
S7.1 Power, organizing
S7.2 Respect
S7.3 Competition
S7.4 Permission
S8 Helping/hindering
S9 Religion and the supernatural

T TIME
T1 Time
T1.1 Time: General
T1.1.1 Time: General: Past
T1.1.2 Time: General: Present; simultaneous
T1.1.3 Time: General: Future
T1.2 Time: Momentary
T1.3 Time: Period
T2 Time: Beginning and ending
T3 Time: Old, new and young; age
T4 Time: Early/late

W THE WORLD & OUR ENVIRONMENT
W1 The universe
W2 Light
W3 Geographical terms
W4 Weather
W5 Green issues

X PSYCHOLOGICAL ACTIONS, STATES & PROCESSES
X1 General
X2 Mental actions and processes
X2.1 Thought, belief
X2.2 Knowledge
X2.3 Learn
X2.4 Investigate, examine, test, search
X2.5 Understand
X2.6 Expect
X3 Sensory
X3.1 Sensory: Taste
X3.2 Sensory: Sound
X3.3 Sensory: Touch
X3.4 Sensory: Sight
X3.5 Sensory: Smell
X4 Mental object
X4.1 Mental object: Conceptual object
X4.2 Mental object: Means, method
X5 Attention
X5.1 Attention
X5.2 Interest/boredom/excited/energetic
X6 Deciding
X7 Wanting; planning; choosing
X8 Trying
X9 Ability
X9.1 Ability: Ability, intelligence
X9.2 Ability: Success and failure

Y SCIENCE & TECHNOLOGY
Y1 Science and technology in general
Y2 Information technology and computing

Z NAMES & GRAMMATICAL WORDS
Z0 Unmatched proper noun
Z1 Personal names
Z2 Geographical names
Z3 Other proper names
Z4 Discourse Bin
Z5 Grammatical bin
Z6 Negative
Z7 If
Z8 Pronouns etc.
Z9 Trash can
Z99 Unmatched

Reproduced from http://ucrel.lancs.ac.uk/usas/ (last accessed 15.09.12).

Appendix II

Detailed word counts for the *Shakespearean Drama Corpus* (*SDC*) and the *Non-Shakespearean Early Modern English Drama Corpus* (*NDC*)

	SDC	*NDC*
Breakdown by sex and genre		
Female comedy characters	77,931	66,020
Female history characters	28,416	36,888
Female tragedy characters	33,880	50,525
Male tragedy characters	252,477	251,713
Male history characters	207,327	203,158
Male tragedy characters	196,515	185,268
Breakdown by sex and date		
Female characters pre-1600	71,159	77,467
Female characters post-1600	69,068	75,966
Male characters pre-1600	357,754	340,380
Male characters post-1600	298,565	299,759
All unknown sex pre-1600	459	2,277
All unknown sex post-1600	0	98
All both sex pre-1600	49	285
All both sex post-1600	0	350
Breakdown by date and genre		
Comedy characters pre-1600	152,370	151,250
Comedy characters post-1600	178,246	167,731
History characters pre-1600	212,610	185,682
History characters post-1600	23,161	54,607
Tragedy characters pre-1600	64,441	83,477
Tragedy characters post-1600	166,226	153,835
Breakdown by sex, date and genre		
Female pre-1600 comedy	35705	34,432
Female pre-1600 history	24547	27,134
Female pre-1600 tragedy	10907	15,901
Male pre-1600 comedy	116457	115,706
Male pre-1600 history	188035	158,379
Male pre-1600 tragedy	53262	66,295
Unknown sex pre-1600 comedy	199	1,000
Unknown sex pre-1600 history	0	0
Unknown sex pre-1600 tragedy	260	1,277
Both sexes pre-1600 comedy	9	112
Both sexes pre-1600 history	28	169
Both sexes pre-1600 tragedy	12	4
Female post-1600 comedy	42226	31588
Female post-1600 history	3869	9754
Female post-1600 tragedy	22973	34624
Male post-1600 comedy	136020	136007
Male post-1600 history	19292	44779
Male post-1600 tragedy	143253	118973
Unknown sex post-1600 comedy	0	17
Unknown sex post-1600 history	0	36
Unknown sex post-1600 tragedy	0	45
Both sexes post-1600 comedy	0	119
Both sexes post-1600 history	0	38
Both sexes post-1600 tragedy	0	193
Word counts from Scott's (1999) *WordSmith Tools* V.3.0		

Appendix III
Stop list used to obtain raw frequencies of content words only

a
about
actually
after
almost
already
also
although
always
am
an
and
another
any
anyone
anything
are
as
at
be
because
been
being
better
between
both
but
by
cannot
canst
could
couldst
dare
did
didst
do
does
done
dost
doth
during
durst
each
enough
even
eventually
every
everyone
for
forward
from
get
gets
go
goes
gone
got
had
hadst
has
hast
hath
have
having
he
her
here
herself
him
himself
his
how
i
if
in
into
is
it
like
made
make
making
many
may
me
more
most
must
my
myself
neither
no
not
nothing
of
often
on
one
or
other
our
out
over
own
seems
shall
she
should
since
so
some
someone
something
such
than
that
the
thee
their
them
then
there
these
they
thine
thing
this
thou
though
through
thy
thyself
to
toward
towards
up
upon
us
usually
very
was
we
were
what
when
where
which
while
who
why
with
without
would
wouldst
ye
yet
you
your

Appendix IV
Word counts of plays in the corpora before and after spelling regularisation

Word counts are from Scott's (1999) *WordSmith Tools* V.3.0. Note that these add up to slightly higher totals than those in Appendix II, where breakdowns by sex of character are totalled. All dialogic text appears to be picked up in both counting processes, and this slight anomaly could not be explained.

SHAKESPEAREAN DRAMA CORPUS (SDC)					
Author	Play title	Play-text ID	Word count before spelling regularisation	Word count after spelling regularisation	Word count difference
Shakespeare	*The Comedy of Errors*	SCCOMERR	14,442	14,425	-17
Shakespeare	*The Taming of the Shrew*	SCSHREW	20,521	20,519	-2
Shakespeare	*Two Gentlemen of Verona*	SCTWOGEN	16,952	16,959	7
Shakespeare	*Love's Labour's Lost*	SCLLL	21,051	21,047	-4
Shakespeare	*A Midsummer Night's Dream*	SCAMIDS	16,210	16,204	-6
Shakespeare	*The Merchant of Venice*	SCMOV	20,998	20,984	-14
Shakespeare	*The Merry Wives of Windsor*	SCMWW	21,343	21,342	-1
Shakespeare	*Much Ado About Nothing*	SCMUCHADO	20,903	20,880	-23
Shakespeare	*As You Like It*	SCAYLI	21,335	21,329	-6
Shakespeare	*Troilus and Cressida*	SCTANDC	25,660	25,645	-15
Shakespeare	*Twelfth Night*	SCTWELFTH	19,509	19,543	34
Shakespeare	*All's Well That Ends Well*	SCALLSWELL	22,626	22,657	31
Shakespeare	*Measure for Measure*	SCMFORM	21,375	21,378	3
Shakespeare	*Cymbeline*	SCCYMB	26,986	26,976	-10
Shakespeare	*The Winter's Tale*	SCWTALE	24,705	24,787	82
Shakespeare	*The Tempest*	SCTEMP	16,134	16,160	26
Shakespeare	*Henry the Sixth Part One*	SHHENVI1	20,608	20,605	-3
Shakespeare	*Henry the Sixth Part Two*	SHHENVI2	24,567	24,567	0
Shakespeare	*Henry the Sixth Part Three*	SHHENVI3	23,392	23,391	-1
Shakespeare	*Richard the Third*	SHRICHIII	28,446	28,412	-34
Shakespeare	*King John*	SHKJ	20,503	20,501	-2
Shakespeare	*Richard the Second*	SHRICHII	21,890	21,883	-7
Shakespeare	*Henry the Fourth Part One*	SHHENIV1	24,107	24,081	-26
Shakespeare	*Henry the Fourth Part Two*	SHHENIV2	25,845	25,840	-5
Shakespeare	*Henry the Fifth*	SHHENV	25,726	25,718	-8
Shakespeare	*Henry the Eighth*	SHHENVIII	23,469	23,504	35
Shakespeare	*Titus Andronicus*	STTITUS	19,908	19,905	-3
Shakespeare	*Romeo and Juliet*	STRANDJ	24,104	24,075	-29
Shakespeare	*Julius Caesar*	STJC	19,208	19,170	-38
Shakespeare	*Hamlet*	STHAMLET	29,829	29,835	6
Shakespeare	*Othello*	STOTH	26,066	26,036	-30
Shakespeare	*King Lear*	STLEAR	25,400	25,395	-5
Shakespeare	*Macbeth*	STMACB	16,586	16,571	-15
Shakespeare	*Antony and Cleopatra*	STANTCLEO	23,914	23,904	-10
Shakespeare	*Coriolanus*	STCORIO	26,694	26,719	25
Shakespeare	*Timon of Athens*	STTIMON	17,842	17,859	17
Total			**798,854**	**798,806**	**-48**

NON-SHAKESPEAREAN EARLY MODERN ENGLISH DRAMA CORPUS (NDC)					
Author	Play title	Play-text ID	Word count before spelling regularisation	Word count after spelling regularisation	Difference in word count
Lyly	*Alexander and Campaspe*	NCALEX	14,635	14,693	58
Lyly	*Gallathea*	NCGALL	14,669	14,631	-38
Greene	*Friar Bacon and Friar Bungay*	NCFRIAR	15,868	15,885	17
Peele	*The Old Wives Tale*	NCOLDWI	7,622	7,584	-38
Chapman	*The Blind Beggar of Alexandria*	NCBLIND	13,072	13,083	11
Heywood	*The Fair Maid of the West Part I*	NCFAIRWEST	16,988	17,074	86
Chapman	*An Humerous Dayes Myrth*	NCANHUM	16,219	16,215	-4
Porter	*The Two Angry Women of Abington*	NCTWOANG	27,546	27,653	107
Anon	*Mucedorus*	NCMUCED	10,859	10,843	-16
Dekker	*Old Fortunatas*	NCOLDFORT	25,964	26,050	86
Heywood	*How a Man May Chuse*	NCCHUSE	21,325	21,324	-1
Jonson	*Volpone*	NCVOLP	29,022	29,185	163
Beaumont & Fletcher	*The Woman Hater*	NCHATER	23,859	23,887	28
Wilkins	*The Miseries of Inforst Marriage*	NCMISER	23,717	23,686	-31
Fletcher	*The Faithful Shepherdess*	NCFAITH	22,035	22,048	13
Jonson	*Bartholomew Fayre*	NCBFAIR	36,764	37,234	470
Massinger	*The Bondman*	NCBOND	21,539	21,608	69
Greene	*The Scottish History of James the Fourth*	NHJAMES	20,048	19,949	-99
Marlowe	*Tamburlaine Part I*	NHTAMI	16,215	16,264	49
Marlowe	*Edward II*	NHEDII	20,615	20,641	26
Peele	*The Famous Chronicle of Edward I*	NHEDI	21,383	21,392	9
Marlowe	*The Massacre at Paris*	NHPARIS	9,750	9,778	28
Peele	*The Battle of Alcazar*	NHALCAZ	10,889	11,682	793
Munday	*The Death of Robert Earl of Huntingdon*	NHDEATH	22,362	22,351	-11
Heywood	*Edward IV Part I*	NHEDIVI	22,502	22,518	16
Heywood	*Edward IV Part II*	NHEDIVII	24,134	24,106	-28
Anon	*The Life of Sir John Oldcastle*	NHOLDC	22,782	22,726	-56
Heywood	*If You Know Not Me, You Know Nobody Part I*	NHIFYOUI	11,401	11,381	-20
Dekker	*Sir Thomas Wyatt*	NHWYATT	10,931	10,948	17
Armin	*The Valiant Welshman*	NHWELSH	17,237	17,256	19
Drue	*The Duchess of Suffolk*	NHDUCH	16,319	16,347	28
Kyd	*The Spanish Tragedy*	NTSPANT	20,857	20,821	-36
Marlowe	*The Jew of Malta*	NTJEW	19,712	19,760	48
Anon	*Arden of Feversham*	NTAOF	19,752	19,735	-17
Marlowe	*Dr Faustus*	NTDRFAUST	11,487	11,505	18
Marlowe	*Dido Queen of Carthage*	NTDIDO	13,608	13,602	-6
Heywood	*A Woman Killed With Kindness*	NTAWKK	16,169	16,184	15
Jonson	*Sejanus*	NTSEJAN	27,614	27,692	78
Beaumont & Fletcher	*The Maid's Tragedy*	NTMAIDSTR	21,187	21,214	27
Webster	*The White Devil*	NTWHITE	26,094	26,286	192
Webster	*The Duchess of Malfi*	NTDOM	24,238	24,480	242
Middleton & Rowley	*The Changeling*	NTCHANGE	18,036	18,243	207
Middleton	*Women Beware Women*	NTWBW	25,878	26,297	419
Total			**832,903**	**835,841**	**2,938**

Appendix V
PHP scripts used for corpus annotation (written by Andrew Hardie)

First PHP script

```
<?php

$file_to_use = $argv[1];

list ($text_id) = explode('.', $file_to_use);

$target_filename = $text_id . ".xml";

$file = file_get_contents($file_to_use);

$file = "<text id=\"$text_id\">\r\n" . $file . "\r\n</text>";

echo strlen($file) . "\r\n";

$file = preg_replace("/\r\n(\w+)\.[ \t]*\r\n/", "\r\n</u>\r\n$1.\r\n<u who=\"$1\">", $file);
//$file = preg_replace("/d/", "D", $file);

$file = preg_replace('/View document image \[\d+\]/', '', $file);

$file = preg_replace('/View this entire document as:\s*‹‹ Back to results/', '', $file);

echo preg_last_error() . "\r\n";

echo strlen($file) . "\r\n";

file_put_contents($target_filename, $file);

?>
```

Second PHP script

```
<?php

// assume that the first thing after the name of the script is the file we want to analyse
$file = $argv[1];

$data = file_get_contents($file);

preg_match_all('/<u who="([^"]*)">/', $data, $matches, PREG_PATTERN_ORDER);

$results = array_unique($matches[1]);

sort($results);

echo implode("\r\n", $results);

?>
```

Third PHP script

```
<?php

$file = $argv[1];

$data = file_get_contents($file);

$replace = 'Rhesus';

foreach ( array('Rhe', 'Rh') as $search  )
{
        $data = preg_replace("/<u who=\"$search\">/", "<u who=\"$replace\">", $data);
}

file_put_contents($file, $data);

?>
```

Fourth PHP script

```
<?php
$total_wc = 0;

$file = $argv[1];

$data = file_get_contents($file);
preg_match_all('/(<u who="[^"]*">)(.*?)<\/u>/s', $data, $matches, PREG_SET_ORDER);

$file_wc = 0;
$people = array();
foreach ($matches as $m)
{
        $file_wc += ($wc = str_word_count($m[2]));
        //echo $m[1] . " word count is " . $wc . "\r\n";
        if (array_key_exists($m[1], $people))
                $people[$m[1]] += $wc;
        else
                $people[$m[1] = $wc;
}

foreach( $people as $person => $words)
        echo "$person\tWords: $words\r\n";

echo"\r\nWord count:\t$file_wc\r\n";

?>
```

Fifth PHP script

```
<?php

$file = $argv[1];
$data = file_get_contents($file);

/* we need to know the text_id so as to insert it into the output */
preg_match('/<text id="([^"]*)">/', $data, $m);
$text_id = $m[1];

/* I can't remember if we've discussed arrays; if not, an array is a bundle of variables that
  you can cycle through, created as follows. */
$array_of_characters = array (
                                                        'Winwife',
                                                        'John Littlewit'
                                                        /* add as many others to this list as you
want */
                                                        /* you would need to amend this
manually at present... for the ultimate
                                                          solution of the prblem, you'd want a
stem for passing this info into
                                                          the script without amending the actual
script */
                                                        );


/* create an empty string to put the results in */
$result = '';


/* a foreach does something to each member of the array in turn. Inside the loop,
  the currently-being-used member of $array_of_characters is accessed using the
  variable $character */
foreach ($array_of_characters as $character)
{
        /* extract all the utterances */
        preg_match_all("/<u who=\"$character\">.*?<\/u>/s", $data, $matches,
PREG_PATTERN_ORDER);
        /* then cycle through each utterance. Change each u tag to add a "from" element before
          adding it to the result string (With a line break after it). */
        foreach($matches[0] as $m)
                $result .= str_replace("<u who", "<u from=\"$text_id\" who", $m) . "\r\n";
}
file_put_contents("$text_id-output-utterances.txt", $result);

?>
```