

A new kind of dictionary for Shakespeare's plays: an immodest proposal

Jonathan CULPEPER
Lancaster University

ABSTRACT

The language of Shakespeare's plays has received substantial treatment in various 'dictionaries', 'glossaries', 'lexica' and 'concordances'. However, the classic works are written in the philological tradition that characterised the *Oxford English Dictionary*. This paper explores how modern principles and techniques developed in Corpus Linguistics can be deployed in the creation of a radically new kind of dictionary. In particular, this involves a focus on usage and frequency. A further innovation is that the proposed dictionary will be comparative, making both internal comparisons (e.g. female characters compared with male) and external comparisons (e.g. Shakespeare's usage compared with that of contemporary plays and other genres). The bulk of this paper is made up of case studies, involving discussion of the words 'horrid', 'good', 'ah' and 'and', multiword units, and linguistic profiles for characters and plays. Through these, the aim is to demonstrate the characteristics of the dictionary and raise pertinent issues, including, for example, how many and what kind of words to include in the dictionary, whether the dictionary should include only words (and how they should be defined), how word-senses should be distinguished, how stylistic and social meanings should be captured, and what approach to grammar should be taken.

KEYWORDS: corpus linguistics, dictionary, Shakespeare, stylistics, words

1. Introduction

The best-known classic Shakespearean 'dictionary' is probably Charles T. Onions's *Glossary* ([1911] 1986), written in the philological tradition that characterised the *Oxford English Dictionary* (OED), and providing pithy definitions and illustrative quotations.¹ The proposed dictionary of the language of Shakespeare's plays is

¹ Onions was in fact one of the editorial team of the OED.

analogous to more recent developments in dictionaries of general English, and, more specifically, the departure from the philological tradition brought about by the *Collins Cobuild Dictionary of the English Language* (Sinclair 1987). The *Collins Cobuild* is a corpus-based dictionary. This implies both a particular methodology for revealing meanings and a particular theoretical approach to meaning, as we shall see in this paper. In particular, there is a strong empirical emphasis. There is less reliance on the vagaries and biases of editors, and a greater focus on the evidence of *usage*. The question of 'what does X mean?' is pursued through another question: 'how is X used?'. To answer the 'how' question, corpus approaches deploy the whole gamut of computational techniques, in order to reveal patterns of usage in context. This inevitably involves matters of *frequency*. Frequency is not in fact as alien as it might seem to the literary critical ear. Any textual analysis that identifies a pattern implicitly involves frequency, as a pattern is the (full or partial) repetition of elements. In fact, the proposed dictionary goes beyond what one might find in the *Collins Cobuild* in a number of ways. Crucially, an additional feature proposed for the dictionary that makes it like no other is that it aims to be *comparative*.² Saying that X word occurs Y times in Shakespeare's plays and that it has W and Z senses is less informative than contrasting those facts with those of his contemporaries (and not just writers of literary texts but writers of various text-types, including records of spoken interaction). In this way, we can reveal not just the denotative or conceptual meanings of words but also their stylistic, discursive and pragmatic values in the general language of the period. Similarly, the plan for the dictionary is that it should also conduct internal comparisons, taking account of the distribution of items over internal genres (e.g. comedy, tragedy, history, particular characters, particular plays) and social categories (e.g. gender, role). Of course, what is revealed through these internal comparisons can be further pursued through external comparisons. For example, having identified that X is typical of women in Shakespeare, one could examine whether X is typical of women in plays by other contemporary playwrights, in 'real life' trial proceedings, and so on.

In this paper, I will deploy a number of case studies to show how techniques developed in corpus linguistics can be used to

² This kind of approach is akin to the corpus-based grammar produced by Biber et al. (1999).

produce the new kind of dictionary based on usage and frequency that I wish to propose. The case studies below are chosen to illustrate particular issues relating to the dictionary; each case study is not complete in itself.

2. Labels and contents of current general Shakespearean 'dictionaries'

I refer to *general* Shakespearean 'dictionaries' in order to exclude 'dictionaries' focusing on specific registers, such as legal, military or informal language (see the *Athlone Shakespeare Dictionary Series*). However, even with this exclusion, identifying what might count as a general Shakespearean dictionary is far from easy. We find various labels for books with contents characteristic – at least to some degree – of dictionaries, notably, 'dictionary', 'glossary', 'lexicon' and 'word-book'. To these one might wish to add 'concordances', in recognition of the fact that such works contain a complete word list and (statistical) information about those words – aspects that might characterise a dictionary. Moreover, what these works contain varies greatly. It is possible to identify three groups. One is strongly *linguistic* in content, typically containing information about the existence of a word-form, as well as its meaning (conveyed with a brief definition and illustrative quotation(s)) and part-of-speech (e.g. Foster 1908, Schmidt [1902] 1971, Onions 1986, Crystal and Crystal 2002). Another group is strongly *non-linguistic* in content, typically containing play summaries (largely plot focussed), character descriptions, cultural information and biographical information (e.g. Boyce 1996, Wells 1998). Note that, although non-linguistic, both of the examples cited are entitled 'Dictionary of Shakespeare'. The final group is strongly focused on (*frequency of*) *occurrence* information, typically containing an index of all words (plus textual location) and the frequency of word-forms (absolute and relative) (e.g. Spevack 1968-80, Howard-Hill 1969-72). There is a little slippage between these groups – for example, Schmidt (1971) contains a complete index of words and Crystal and Crystal (2002) was constructed with frequency information in mind – but in the main they are separate. My proposal involves bringing together the three areas in a more comprehensive and systematic fashion.³

³ This will clearly involve a broad scope. Consequently, the label 'Dictionary of Shakespeare' may not be the best. An alternative might be 'Encyclopaedia of Shakespeare's Language' (I am grateful to Anthony Warner for this suggestion).

3. General Shakespearean 'dictionaries' and present-day English Language dictionaries compared

The majority of present-day dictionaries of English contain pronunciation information (typically, a broad phonetic transcription with an indication of syllable stress). Doing the same for Shakespeare would require a significant research programme, and there would be thorny issues, such as whose accent to represent. Consequently, this is not currently part of the dictionary proposal. Many present-day dictionaries contain spelling variants, and the OED, of course, excels in this respect. Shakespearean dictionaries do not note more than the occasional spelling variant. Perhaps spelling variants are assumed not to be part of the 'real' Shakespeare, given that they are produced by compositors and printers. Nevertheless, spellings are the prism through which we receive Shakespeare, and Shakespearean texts represent a source of information about spelling in the early modern period. Moreover, quantifying spelling variation would be relatively easy to do with the computational methodology supporting the proposed dictionary (see section 12). Other differences in content include the fact that corpus-based dictionaries of present-day English, notably, the Collins Cobuild dictionary, include definitions that are more contextualised and information about multi-word units, as I will illustrate in sections 4 and 9.

Perhaps even more significant than differences in the kinds of information that might be included are differences in policies for including or excluding words and for prioritising meanings. Shakespearean 'dictionaries', notably, Foster (1908) and Onions (1986), but even more recent corpus-informed dictionaries such as Crystal and Crystal (2002), tend to include only those words considered difficult or 'hard' for readers. In contrast, corpus-based dictionaries typically include all the words in the corpus (though that may not, in fact, be the best thing to do for a Shakespearean dictionary; see section 5). Furthermore, present-day dictionaries, particularly corpus-based dictionaries, take a different approach to the way meanings are prioritised within particular entries. Dictionaries in the philological tradition exemplified by the OED (e.g. Foster 1908 and Onions 1986) take etymology as a guiding principle. This is most obviously reflected in the way that (1) word definitions gravitate towards etymological meanings, and (2) the organisation of the senses of polysemous words is based on etymological priority (i.e. the earliest sense is listed first). In contrast, corpus-based dictionaries capture meanings based on usage in

context, and organise those meanings according to frequency (usually the most frequent is placed first).

4. Towards a contextualised definition: the case of 'horrid'

The OED gives three senses for the word 'horrid': (1) "bristling, shaggy, rough," (2) "causing horror or aversion; revolting to sight, hearing, or contemplation; terrible, dreadful, frightful; abominable, detestable," and (3) "*colloq.* in weakened sense. Offensive, disagreeable, detested; very bad or objectionable. Noted in *N.E.D.* as especially frequent as a feminine term of strong aversion" (here, and in all quotations from dictionaries in this paper, accompanying quotations are generally excluded for brevity). The first sense corresponds with that of the Latin term 'horridus' from which the English word is derived, and, judging from the illustrative quotations, was still current in Shakespeare's period. The second sense, and one that is contemporary with Shakespeare, is a metonymic development of the first, and the final sense is apparently a 'weakened' development of the second. The fact that the first quotation given to illustrate the second sense is from Shakespeare should alert us to a major problem in using the OED to interpret Shakespeare – the problem of circularity, given that Shakespeare plays such a large role in determining the entries in the OED for the period in question. The third sense developed after Shakespeare. Note that the OED does at least supply a modicum of stylistic information, noting that the third sense is colloquial, and very occasionally some social information, here noting that the third sense is "especially frequent as a feminine term."⁴

Turning to three Shakespearean dictionaries, we find the following definitions:

Foster (1908): (1) Awful, hideous, horrible. (2) Terrific.⁵ (3) Horrified, affrighted.

Onions (1986): No entry.

Crystal and Crystal (2002): Horrifying, frightful, terrifying.

Foster's (1908) first definition seems to shade into the third sense given in the OED. This is odd because the first citation date for that

⁴ Although the evidence is thin, explorations in the present-day *British National Corpus* suggest that women do tend to use the term 'horrid' more than men.

⁵ In the period Foster was writing, this could have the earlier sense of 'causing terror.'

sense given in the OED is 1666. The single illustrative quotation given by Foster is from Macbeth: 'If good, why do I yield to that suggestion / Whose *horrid* image doth unfix my hair.' This quotation includes a classic reaction to fear – the unfixing of hair. This does not support the sense given by Foster, which need not involve fear, just as in the third sense given in the OED. In fact, the usage here falls within the scope of the OED's second sense, as indeed do Foster's second and third definitions.

Note that the strongly overlapping array of synonyms given in the definitions do little to pin down the sense of 'horrid' in Shakespeare. What is being described as horrid? Who is using this word? In what circumstances are they using it? Is Shakespeare using it in a way that his contemporaries would not? And so on. We can look at a computer concordance (a list of the occurrences of the word along with their local co-text) and the distribution of a word, in order to answer such questions. Here is the entire concordance of 'horrid' (the head noun to which it refers is underlined):⁶

Appear in forms more **horrid**, – yet my duty, As doth a Rock
 Up Sword; and know thou a more **horrid hent**: When he is drunk asleep⁷
 And cleave the general ear with **horrid speech**, Make mad the guilty
 heard and seen, Recounts most **horrid sights** seen by the watch
 shall break his wind With fear and **horrid flight**. 1.Sen. Noble
 I will meditate the while upon some **horrid message** for a Challenge.
 Macd. Not in the legions of **horrid hell**, can come a devil more damned
 Proper deformity seems not in the fiend So **horrid** as in woman.
 And what a beard of the general's cut and a **horrid suit** of the camp
 Presented then unto the gazing moon So many **horrid ghosts**.
 Crammed with distressful bread; Never sees **horrid night**, the child of hell
 all the sparks of nature, To quit this **horrid act**. Reg. Out treacherous
 Such sheets of fire, such bursts of **horrid thunder**, Such groans of
 couriers of the air, Shall blow the **horrid deed** in every eye,
 Of thy dear husband, than that **horrid act** Of the divorce he'ld make
 I yield to that suggestion, Whose **horrid image** doth unfix my hair

It can be seen that 'horrid' is used to describe acts, sights and sounds, but not just any such things – most have a strong

⁶ A concordance of a word will vary in terms of how many instances it contains according to the edition of Shakespeare used (and occasionally according to how good the search software is). The particular Shakespeare edition used in this paper is outlined in footnote 8.

⁷ 'Hent' means 'clasp'.

supernatural connection. This seems to have been overlooked in all dictionary definitions, despite the fact that it is quite obvious in the concordance. We can deepen our understanding of the word by considering its distribution both within Shakespeare and without. Putting the results together, a dictionary entry might be as follows (All = All Shakspeare's plays, T = tragedies, C = comedies, H = histories, M = male speakers, F = female speakers, Pla = other EModE plays, Fic = EModE prose fiction, Tr = EModE trial proceedings, Ha = EModE handbooks in dialogue form, Sc = EModE scholarly works; the figures in brackets are normalised per 100,000 words):

HORRID. Something that is **horrid** causes fear; typically, it refers to supernatural or unnatural acts, sights and sounds. Distribution: All = 16 (1.8); T = 10 (3.9), C = 2 (0.6), H = 4 (1.5); M = 14 (1.9), F = 2 (1.4). Comparisons: Pla = 187 (0.17), Fic = 0, Tr = 0, Ha = 0, Sc = 1 (0.14). EG 'Whose **horrid** Image doth vnfixe my Heire', 'I wil meditate the while vpon some **horrid** message for a Challenge'.⁸

The above is no more than an indication as to the direction a dictionary entry might take. Note that the first sentence offers a contextualised definition of the type used in the *Collins Cobuild*, rather than a handful of synonyms. However, going beyond the *Collins Cobuild*, the figures following offer a broader discursal contextualisation. They give some indication as to the social and stylistic meanings the word might have acquired on account of being to some degree 'contextually bound' (Leech 1981: 14-15; see also

⁸ The Shakespeare frequencies given in this paper are based on *The Nameless Shakespeare* (2003), a joint project of the Perseus Project at Tufts University, The Northwestern University Library, and Northwestern University Academic Technologies. It is derived from The Globe Shakespeare, the one-volume version of the Cambridge Shakespeare, edited by William George Clark, John Glover and William Aldis Wright (1891-3). There is no claim here that this constitutes the ideal edition of Shakespeare. It is searchable via 'WordHoard' (the concordance in section 4 was derived by this). The comparative 'Pla' corpus is the 'Korpus of Early Modern Playtexts in English' (KEMPE), initially compiled by Lene B. Petersen and Marcus X. Dahl, University of Bristol, 2001-2003. It is searchable via 'Corpuseye'. Note: a particular problem with the Corpuseye search engine is that it only searches the whole corpus and that corpus includes Shakespeare. Nevertheless, given the great size of the corpus – 10.7 million words – the results will still mean something. The samples for early modern prose fiction, trial proceedings and handbooks are sourced from the *Corpus of English Dialogues 1560-1760*, and the scholarly works comprise half history writing and half science writing, sourced from the *Helsinki Corpus of English Texts*.

Enkvist 1964: 29-35). Focusing on the more meaningful normalised figures, one might note, for example, that the word 'horrid' appears much more densely in tragedies than either histories or comedies, is used slightly more frequently by male characters compared with female; and that Shakespeare uses it considerably more than his contemporary playwrights did, and also that it is most characteristic of Early Modern plays and, surprisingly, scholarly literature.

However, this particular example is severely hampered by frequency limitations: the strongest finding revealed by the figures simply being that 'horrid' is rare. I will focus on frequency limitations in the next section. Here, I will briefly indicate four ways in which the above entry could be improved:

- The definition was derived from collocational information and some of this information could have been included in the entry (see section 6).
- Sociolinguistic information could be enriched by the inclusion of other sociological variables (e.g. status, age) and also comparative data (e.g. addressing questions such as: is X word associated with male or high status speakers in Shakespeare specifically or is this a more general feature of Early Modern English?).
- A statistical measure could be employed in order to indicate whether differences in distribution are significant.
- The presentation of information could be improved (e.g. the use of graphs, or a verbal description instead of figures).

5. Frequency limitations

A corpus-based dictionary typically includes all words in the corpus. However, this presents two problems: (1) how to treat rare or infrequent words, and, from the more practical point of view of publication, (2) how to fit all the words into one volume.⁹ As is clear from the sample entry of 'horrid' above, low frequency words lead one to the mere conclusion that they are low frequency, as the more robust and informative distribution patterns fail to materialise.¹⁰ The corpus-based methodology is not best suited to investigating low frequency words (cf. Biber et al. 1998: 30, Meyer 2002: 15), instead we

⁹ Schmidt's (1971) complete treatment of Shakespeare's lexicon stretches over two volumes of small print and thin paper, yet only contains the briefest of definitions.

¹⁰ There are also difficulties in applying statistical significance tests to differences in distribution that involve low frequencies.

need to look towards alternative methodologies, such as the philological approach that already underpins most current Shakespearean dictionaries. A partial solution to these problems is simply to adopt a frequency cut-off point such that words below a certain frequency are not considered for inclusion in the dictionary. But what would be the implications of such a cut-off point for the coverage of Shakespeare's vocabulary?

Onions (1986) supposedly covers some 3,000 words, according to Crystal and Crystal (2002: Introduction), who also claim to include 21,263 entries under 13,626 headwords in their own volume. Table 1 displays the consequences of various cut-off points for the number of different word-forms (types) and for the total number of word-forms (tokens) that appear in Shakespeare.

	Types	Tokens
Word-types with more than 100 instances	998	706,974
Word-types with more than 50 instances	1,564	761,472
Word-types with more than 16 instances	4,652	835,925
Word-types with less than 10 instances	7,753	37,260
All instances of all word-types	24,842	899,092

Table 1. Word-form types, tokens and cut-off points in Shakespeare's plays

As the bottom row shows, there is a total of 899,092 word-tokens in Shakespeare and 24,842 different word-types (in other words, a smaller number of different words are repeated a number of times to make up the total vocabulary of Shakespeare's plays). 'Horrid' occurred 16 times. If we only consider for the proposed dictionary word-types that occur more than 16 times, then, potentially, we would only need to have 4,652 different entries in our dictionary, and yet we would still cover most of the language of Shakespeare's plays (835,925 word-tokens). However, I pointed out above that 16 occurrences is too few for our purposes. If a cut-off point of, say, 50 word-tokens for any entry were imposed, resulting in a potential and certainly more manageable 1,564 word-form entries, then that still would account for the vast bulk of the words in Shakespeare (761,472 out of 899,092 word-tokens). Note that 7,753 word-types occur less than 10 times, accounting for a mere 37,260 word-tokens. Yet it is precisely here that the current Shakespearean dictionaries tend to focus, as these rare items tend to be considered 'hard'. However, in my view, there is no justification for excluding more frequently occurring vocabulary items. From a linguistic perspective,

we know that all words change meaning: even the most frequent of items have incurred shifts of meaning which present-day readers must take on board. From a literary perspective, we should beware of letting more unusual vocabulary distract our attention from the more usual. As John F. Burrows (1987: 1) eloquently puts it: 'It is a truth not generally acknowledged that, in most discussions of works of English fiction, we proceed as if a third, two-fifths, a half of our material were not really there.'

6. Polysemy and collocates: the case of 'good'

Current Shakespearean dictionaries give definitions for the word 'good' (as an adjective) such as these (illustrative quotations are excluded):

Foster (1908): (1) Not bad, worthy of praise; (2) Fit, adapted; (3) Trustworthy, genuine; (4) Kind, benevolent; (5) Proper, right; (6) Substantial, safe, solvent, able to fulfil engagements, (7) Real, serious; (8) Favourable, propitious, (9) Abundant, rich, (10) Skilful, clever, (11) Adequate. Notes phrases and compounds.

Onions (1986): (1) Conventional epithet to titles of high rank, (2) comely, (3) Financially sound; (hence) wealthy, substantial. Notes quasi-adverbial usage, e.g. 'good easy man', and phrases and compounds

Crystal and Crystal (2002): (1) [intensifying use] real, genuine ('love no man in good earnest'). (2) kind, benevolent, generous. (3) kind, friendly, sympathetic. (4) amenable, tractable, manageable. (5) honest, virtuous, honourable. (6) seasonable, appropriate, proper. (7) just, right, commendable. (8) intended, right, proper. (9) high-ranking, highborn, distinguished. (10) rich, wealthy, substantial. Notes phrases and compounds

Lists of synonyms – in some cases overlapping – do not always provide the reader with assistance in discriminating the various senses. For example, in *Foster's* (1908) definitions, how does 'genuine' in sense 3 differ from 'real' in sense 7? Similarly, 'fit' in sense 2 can uncomfortably overlap with 'proper, right' of sense 5. *Onions's* (1986) definitions are fairly discrete, whilst in contrast *Crystal and Crystal* (2002) seem to have gone for a deliberate policy of overlap (note that 'kind', 'proper' and 'right' appear in more than one definition), perhaps indicating that indeed senses do overlap. We might also note that each dictionary orders the senses in a different way, and that some, rather worryingly, contain senses that others do not (note, for example, *Onion's* first sense).

A simple technique in corpus linguistics for investigating the meaning of a word is to examine a concordance and note the words with which the word in question co-occurs, something which we have already demonstrated with the word 'horrid'. It is the collocates of a word – "the company it keeps" (cf. J.R. Firth 1957) – that may help distinguish different senses (see, for example, Partington 1998: 33-46). Frequent collocating words to the right of 'good' include: '(my) *good* friend(s)/sir/Lord/master/man/Lady/Madam/etc.,' '*good* old man/friend/etc.,' '*good* morrow/night/even,' '(in) *good* faith,' '*good* will/wish(es),' '*good* god(s),' '*good* luck /hap,' '*good* news/report/words,' '*good* now,' and '(in) *good* time.' Even without further elaboration, seeing such collocations helps make accessible distinct senses, and so they should be included within a dictionary entry. Also, the frequency of such collocations can feed into the ordering of senses within the entry. However, with a dizzying 2711 instances constituting a concordance of 'good', the human can only identify some collocational patterns, and cannot accurately assess the strength of those patterns and thus come to a principled decision about which to include in the dictionary. One possible solution is to calculate the statistical likelihood with which particular words and 'good' co-occur to form a collocation. Using z-scores, a statistical measure, the top 10 ranked-ordered collocates five words to the left and right of 'good' are: *morrow*, *Lord*, *my*, *do*, *sir*, *good*, *your*, *have*, *be*, and *you*.¹¹

Collocate (+5/-5)	Frequency	Z-score
<i>Morrow</i>	6	18.0
<i>Lord</i>	11	4.8
<i>My</i>	21	2.8
<i>Do</i>	8	2.5
<i>Sir</i>	6	2.1
<i>Good</i>	6	2.1
<i>Your</i>	11	2.0
<i>Have</i>	9	1.5
<i>Be</i>	9	1.1
<i>You</i>	15	0.7

Table 2. The top 10 ranked-ordered collocates of 'good' within a five-word span

¹¹ It is a matter of debate as to which statistical measure to use. Mutual information scores are frequently used, some use t-scores and some argue for the Fisher exact test. These results were produced using the software *Xiara*.

These collocational patterns point to sentences like the following (constructed) example: 'Good morrow, my good lord, you have [...].' This evidence clearly underpins Onions' (1986) first sense, a sense that is not clearly represented in the other dictionaries, and underscores the role of 'good' as a politeness marker. Such an investigation could be extended in three ways: (1) collocational patterns (and ones not limited to single word collocates) can be identified with other statistical procedures (including the methodology in section 8), (2) collocational patterns in Shakespeare can be compared with collocational patterns in other Early Modern texts (e.g. is Shakespeare peculiar in his usage of 'good' as a politeness marker?), and (3) grammatical relations can be explored via collocations (e.g. as is transparent for the concordance of 'horrid' the items immediately to the right are nouns – something that confirms the status of 'horrid' as a typical adjective).

7. The inclusion of pragmatic/discoursal words: the case of 'ah'

Interjections, onomatopoeic sounds, hesitation phenomena, discourse markers, and so on have received scant attention in Shakespearean dictionaries (of course, this is not true of specialist dictionaries, notably Blake 2004). For example, there is no entry for 'ah' in Foster (1908), Onions (1986) or Crystal and Crystal (2002). The issue is whether such items are considered words, and that depends on your definition of the word. Corpus linguistics favours an orthographic definition, such as 'a string of uninterrupted non-punctuation characters with white space or punctuation at each end' (Leech et al. 2001: 13-14). In which case, 'ah' is clearly a word. Does 'ah' have meaning? That depends on your definition of meaning. If meaning is associated with ideational meaning, to use Halliday's (e.g. 1978) terminology, and not textual or interpersonal meanings, then words like 'ah' do not have meaning. One of the reasons such words are not generally included in Shakespearean dictionaries is that words that reflect some aspect of the world are privileged above words that help organise other words or words that help organise people. In my view, this approach is entirely *inappropriate* for a dictionary of Shakespeare's plays because those plays are made up of *dialogue*. What lies at the heart of dialogue are those pragmatic and discoursal words that structure and mediate the interaction between characters.

Let us consider the pragmatic and discursal meanings of 'ah', and also its social and stylistic meanings. If a concordance of 'ah' is scrutinised, one can discern the three key pragmatic meanings following (an illustrative example is provided of each):

- (1) *Speaker attitude/state communicated = sorrow, emotional distress*
Des. To whom my Lord? With whom? How am I false?
Oth. **Ah** Desdemona, away, away, away.
Des. Alas the heavy day: why do you weep? Am I the motive of these tears my Lord? *Othello*
- (2) *Speaker attitude/state communicated =pity*
Glou. Canst thou blame him? His daughters seek his death: **Ah**, that good Kent, He said it would be thus: poor banish'd man: Thou sayest the King grows mad, I'll tell thee friend I am almost mad my self.
King Lear
- (3) *Speaker attitude/state communicated = surprise, realisation*
 [Enter Adriana and Luciana.]
Adr. **Ah** Luciana, did he tempt thee so? *Comedy of Errors*

And one can discern the two key discursal meanings following:

- (1) *Discourse marker: preface to the correction / rejection of the previous speaker's proposition(s), emotions or actions*
Men. These three world-sharers, these competitors are in thy vessel. Let me cut the cable, And when we are put off, fall to their throates: All there is thine.
Pom. **Ah**, this thou shouldst have done, And not have spoke on't. In me 'tis villany, In thee, 't had bin good service: [...] *Antony and Cleopatra*
- (2) *Discourse marker: reinforces elicitation*
Leon. All thy tediousnesse on me, **ah**?
Const.Dog. Yea, and 'twere a thousand times more than tis, for I hear as good exclamation on your Worship as of any man in the City, and though I be but a poor man, I am glad to hear it. *Much Ado about Nothing*

Turning to stylistic and social meanings, consider the distribution of 'ah':

Distribution: All = 179 (19.9); T = 54 (21.3), C = 32 (8.9), H = 93 (35.4); M = 121 (16.1), F = 59 (41.9). Comparisons: Pla = 1573 (14.4), Fic = 9 (10.9), Tr = 1 (2.9), Ha = 11 (11.2), Sc = 0.

Within Shakespeare, 'ah' is characteristic of the histories, to some extent the tragedies, but to a much lesser extent of comedies. This distribution may reflect the frequent functions of 'ah' in signalling emotional distress and pity. Interestingly, the distribution across genders is far from even: it is more than twice as dense in female dialogue. Compared with other playwrights of the period, Shakespeare can be said to be fairly fond of this item. Also, we can see that it is more characteristic of plays than other contemporary genres. Moreover, there is evidence that 'ah' is a strong colloquial marker. It does not appear at all in scholarly works, the genre that is far removed from colloquial genres; it hardly appears in trial proceedings, a genre that – influenced by the formal setting, legal routines and need to create an official document – tends to be remote from colloquial language; whilst on the other hand, it appears in fictional prose (the choice of prose for this dataset being specifically geared towards more colloquial prose) and handbooks in dialogue form. Interestingly, and remarkably, the density of 'ah' in a sample of five present-day plays is 94.27 (contrasting with Shakespeare's 19.9), something which presumably reflects the drift of genres, including plays, towards more colloquial language (see, for example, Biber and Finegan 1992).

8. The inclusion of grammatical words: the case of 'and'

The most frequent words in any body of texts are closed-class. Yet Shakespearean dictionaries do not, or do not adequately, treat such grammatical items, despite – or may be because of – their high frequency of occurrence. For example, the entries for the second most frequent word in Shakespeare, the word 'and', in general Shakespearean dictionaries are as follows:

Foster (1908): Cross-references Abbot's Shakespearean grammar.

Onions (1986): (1) Coordinating conjunction (nouns, adjectives and phrases);

(2) Subordinating conjunction: if, even if, though, as if, whether.

Crystal and Crystal (2002): [also spelling variant 'an'] (1) if, even if; (2) as if;

(3) if, whether.

As can be seen, it is not treated at all in *Foster* (1908), whilst *Crystal and Crystal* (2002) only mention *conditional* 'and' (used as a subordinate conjunction introducing a conditional clause with the sense 'if'). Conditional 'and' is likely to be the focus of attention in Shakespearean dictionaries, because of editorial policies to select

items with which the modern reader is assumed to be unfamiliar and thus likely to experience difficulty. Examples of conditional ‘and’ include the following:

What would you have me be, **and** I be not a woman? *Pericles*

Noting this penury, to my self I said, **An** if a man did need a poison now, Whose sale is present death in Mantua, Here lives a caitiff wretch would sell it him. *Romeo and Juliet*

Only Onions (1986) acknowledges the fact that words such as ‘and’ play an important grammatical role. It is the coordinating function of ‘and’ that accounts for the overwhelming majority of instances in Shakespeare. ‘And’ makes a significant contribution to textual meaning in Shakespeare in the way it conjoins nouns, adjectives, (nominal or adjectival) phrases and clauses, and it is also used as a pragmatic connective.

Compare the following two extracts in which instances coordinating clauses are underlined and instances coordinating words/phrases are emboldened:

Duke. She should this Angelo have married : was affianced to her oath, and the nuptial appointed: between which time of the contract, **and** limit of the solemnity, her brother Frederick was wrecked at Sea, having in that perished vessel, the dowry of his sister: but mark how heavily this befell to the poor Gentlewoman, there she lost a noble **and** renowned brother, in his love toward her, ever most kind **and** natural: with him the portion **and** sinew of her fortune, her marriage dowry: with both, her combinate-husband, this well-seeming Angelo. *Measure for Measure*

Citizen2 [...] Who’s that that bears the Sceptre?

Citizen1 Marquess Dorset, And that the Earl of Surrey, with the Rod.

Citizen2 A bold brave Gentleman. That should be The Duke of Suffolk.

Citizen1 ‘Tis the same: high Steward.

Citizen2 And that my Lord of Norfolk?

Citizen1 Yes. *King Henry VIII*

The density of lexical/phrasal coordination in the first extract contrasts with the density of clausal coordination in the second. This grammatical difference reflects differences in style and communicative purpose. Lexical/phrasal coordination in the Dukes

speech helps create a high rhetorical style, underscoring the seriousness of what he is saying. More specifically, the conjoins of coordinated pairs tend to be closely related in meaning. Thus, 'noble' and 'renowned' overlap in meaning (reflecting the rhetorical figure of 'pleonasm'), and 'kind' and 'natural' could be viewed as being in a hierarchical relationship such that one is subordinate to the other (reflecting the rhetorical figure of 'hendiadys', i.e. amounting to: 'naturally kind'). In contrast, the clausal coordination of the second extract creates a low rhetorical style, underscoring the casual conversation, a style which is, of course, reinforced by the ellipsis. In fact, in this particular case, 'and' is not merely coordinating clauses but also acting as a pragmatic connective. Specifically, it is used to create a series of questions, or, as Schiffrin puts it, to "link questions in a question agenda" (1994: 146). As a consequence of their rather different functions, lexical/phrasal coordination tends to correlate with rather different genres compared to clausal coordination. I cannot prove this claim with regard to Shakespeare, as the computational analysis of Shakespeare's grammar is not yet sufficiently accurate or sophisticated; indeed, one of the aims of my dictionary project is to solve this (see Culpeper and Kytö 2002, which provides supporting evidence for four Early Modern genres). In sum, my argument is that such grammatical items should be included in a dictionary of Shakespeare, and that dictionary should focus widely on the contribution of those items to meaning.

9. Multiword units

John Sinclair (e.g. 1991), amongst other linguists, has argued that words may belong to semi-fixed phrases that constitute single lexical choices (e.g. 'of course', where the individual words cannot be assumed to produce the sense of the phrase). Current Shakespearean dictionaries pay scant attention to these. An empirical way of retrieving lexical items that tend to bunch together is to run an n-gram analysis. 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. The results for Shakespeare, retrieved by *WordSmith Tools* (Scott 1999), are included in Table 3, along with the results for three other datasets for comparison (the underlining, italics and boldening show that a particular lexical bundle is used

in another data set; no lexical bundle is used in more than two data sets).

Shakespeare	EModE Plays	EModE Trials	Present-day Plays
I pray you	<u>it is a</u>	do you know	I don't know
<u>I will not</u>	<i>what do you</i>	I did not	<i>what do you</i>
<u>I know not</u>	<u>and I will</u>	did you see	I don't want
I am a	it is not	I do not	do you think
I am not	I have a	he told me	do you want
my good lord	<u>I will not</u>	at that time	I don't think
there is no	in the world	out of the	to do with
I would not	I tell you	I told him	do you know
<u>it is a</u>	<u>I know not</u>	he did not	going to be
<u>and I will</u>	I warrant you	there was a	don't want to

Table 3. The top ranked-ordered 3-word lexical bundles in Shakespeare and other genres

It has been noted in the literature that lexical bundles are good discriminators of different styles (e.g. Stubbs and Barth 2003). The bulk of the items in Table 3 are unique to the specific data sets. Lexical bundles in Early Modern trials reflect the fact that that discourse is made up of question-answer routines (e.g. 'do you know', 'did you see' versus 'I did not,' 'I do not') and crime-narrative report (e.g. 'he told me,' 'at that time,' 'out of the,' 'I told him'). Lexical bundles in present-day plays seem to gravitate towards questions and assertions to do with knowing, wanting and thinking – perhaps the essence of present-day drama in which plot and character development is conveyed through highly interactive character-to-character dialogue (in other words, what is said between characters is partly designed to inform the audience of character and plot). A characteristic of both Shakespeare and other Early Modern plays is that many of the bundles begin with the first person pronoun 'I', perhaps reflecting the essence of Early Modern drama with its more direct presentation of characters and plot to the audience (the epitome of this being the use of soliloquies and asides). 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, of present-day plays.

The kind of distributional stylistic information I have been discussing here could, of course, be recorded along with the entry for the most frequent lexical bundles in Shakespeare in the dictionary. Perhaps even more importantly, such n-gram analysis can feed into the grammatical description contained in the dictionary. I will attend to this issue in the following section.

10. A note on grammatical description

Linguists like Sinclair (e.g. 1991, 2004) emphasise that grammar is in the lexicon and not in some *a priori* set of abstract categories (e.g. parts of speech) imposed on the language. A way into describing the lexico-grammar of Shakespeare would be to describe the grammatical frames or patterns, revealed through collocational analyses (as discussed in section 6) and multiword analyses (as discussed in section 9) (see Hunston and Francis 2000, for this approach). I have already hinted that collocational analyses could be deployed in the exploration of grammatical relations, noting the case of 'horrid' (and 'good' is similar).¹² In fact, my discussion of 'and' was very much geared towards the grammatical relations of co-occurring units. Regarding multiword units, 'I pray you,' for example, is a grammatical pattern consisting of a first person pronoun (i.e. either 'we' or 'I'), a verb in the present tense and a second person pronoun (i.e. either 'you' or 'thee'). Whilst the items that can occur as pronouns are relatively restricted, a much wider range of verbs can occur in this particular pattern. However, not any verb can occur: the set is restricted. One subset of those verbs is comprised of speech act verbs such as 'advise', 'arrest', 'assure', 'beseech', 'charge', 'tell', 'thank' and 'warrant'. Such verbs occur when the grammatical pattern is used in isolation or parenthetically to a matrix clause. Making the step from an n-gram analysis to the description of grammatical patterns or frames is not necessarily straightforward. N-gram analysis results in units which are not necessarily complete idioms or grammatical structures. Nevertheless, such analysis offers a way into it identifying

¹² As I have already indicated in this paper, a highly accurate part-of-speech tagged corpus of Shakespeare does not exist. Also, there are issues to do with the compatibility of tags and software, as well devising software to assess adequately grammatical relations. One possibility to be explored is *SketchEngine* (see Kilgariff et al. 2004), used for lexicography by Oxford University Press, for example.

grammatical frames, and the results can be complemented by collocational analyses.

I would not argue for quite as radical an approach to grammar (i.e. ditch all abstract grammatical categories) as Sinclair, for four reasons. First, my analysis of 'and' already demonstrated that grammatical categories can be useful. Knowing the grammatical status of the conjoins (i.e. lexical/phrasal versus clausal) helps us account for textual meanings. Second, grammatical categories can provide a useful way of tracking variation and change in the language; specifically in the case of the dictionary, it can help provide a way of understanding how language varied in Shakespeare's time (e.g. from register to register, from person-to-person) and how language has changed since Shakespeare. For example, the proposed dictionary could quantify parts of speech, particularly in cases where an item can function has more than one part of the speech, and thereby reveal differences in distribution (e.g. the distribution of verbal vs. nominal usages of the lexeme 'love' used to be weighted in favour of nominal but is now approximately even). Third, supplying such information about words would enable researchers to compare and contrast with extant research. Fourth, supplying such information can simply be one additional means by which a dictionary can help users understand words.

11. Character and play profiles

Some Shakespearean dictionaries contain non-linguistic descriptions of characters and plot summaries. I propose providing a description of the idiolect of each major character. This can be done by conducting a statistical comparison between the vocabulary of one character and that of the other characters in the same play, in order to reveal words that are statistically characteristic of particular characters. Those words are 'keywords'. As an illustration, consider some of the results relating to characters in *Romeo and Juliet* (see Culpeper 2002, for a more detailed discussion). Table 4 contains the keywords of Romeo and Juliet (rank-ordered in terms of the statistical 'keyness') produced by the program *WordSmith Tools*:

<i>Romeo</i>	<i>Juliet</i>
Beauty (10), Love (46), Blessed (5), Eyes (14), More (26), Mine (14), Dear (13), Rich (7), Me (73), Yonder (5), Farewell (11), Sick (6), Lips (9), Stars (5), Fair (15), Hand (11), Thine (7), Banished (9), Goose (5), That (84)	If (31), Be (59), Or (25), I (138), Sweet (16), My (92), News (9), Thou (71), Night (27), Would (20), Yet (18), That (82), Nurse (20), Name (11), Words (5), Tybalt's (6), Send (7), Husband (7), Swear (5), Where (16), Again (10)

Table 4. Rank-ordered keywords for Romeo and Juliet (raw frequencies in brackets)

This reveals, for example, the predictable result that Romeo's two most unusually frequent words (or 'keywords') are 'beauty' and 'love', but the less predictable – and thus possibly more interesting – result that Juliet's two most unusually frequent words are 'if' and 'be'. Although the results for Juliet are less predictable, they can readily be explained by a qualitative analysis of the text (i.e. they are motivated). Furthermore, and following the line of argument articulated above, although many of Juliet's keywords are grammatical in nature, they are no less meaningful. Upon closer inspection of Juliet's keywords, one can see that keywords such as 'if', 'be' (often subjunctive), 'or', 'would' and 'yet' reflect Juliet anxieties and worries about Romeo's intentions and welfare, as the following examples illustrate:

If he **be** married, / Our grave is like to **be** our wedding-bed (I.v.)
 If they do see thee, they will murder thee (II.ii.)
 But **if** thou meanest not well (II.ii.)
 Is thy news good, **or** bad? answer to that; Say either, and I'll stay the
 circumstance: Let me **be** satisfied, is 't good **or** bad? (II.ii)
 Tis almost morning; I **would** have thee gone; And **yet** no further than
 a wanton's bird [...] (II.ii.)

The key point about such analysis is that, although a reading of the play would obviously have resulted in an understanding of Juliet's anxieties and worries, such a reading would not necessarily have led to the identification of the linguistic source of that very understanding. Indeed, no 'manual' critical analysis to date, literary or linguistic, has accounted for the source.

Regarding plays, plot summaries tend to include information about the plays' 'themes'. Such information relies on the intuitions of the editor. I propose something more empirical: providing a description of the semantic categories (or lexical fields)

characterising each play. This can be done by getting the computer automatically to assign each word in the plays to a semantic category (this assignment can, of course, be recorded in the entry for each word). The dominance of categories within plays can be statistically compared. For example, in an earlier study I conducted with Dawn Archer and Paul Rayson (Archer et al. forthcoming), we compared three 'love tragedies' (*Othello*, *Anthony and Cleopatra* and *Romeo and Juliet*) with three 'love comedies' (*A Midsummer Night's Dream*, *The Two Gentlemen of Verona* and *As You Like It*). Each word was assigned to the categories in Table 5 using the USAS suite of programs (for further details, see section12).

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			

Table 5. The semantic categories used (derived from McArthur 1981)

Then a statistical comparison was conducted in order to establish which semantic categories were characteristic of each data set (each semantic category has several subcategories). Our findings are displayed in Table 6.

<i>Most overused categories in comedies relative to tragedies</i>	<i>The most overused categories in tragedies relative to comedies</i>
S3.2 = intimate/sexual relationship	G3 = warfare, defence, & the army
L2 = living creatures	L1- = (lack of) life/living things
L3 = plants	Z2 = geographical names
S1.2.6- = (not) sensible	E3- = (not) calm/violent/angry
X3.1 = sensory: taste	M4 = movement (by sea/through water)
E2+ = liking	S9 = religion and the supernatural
T3- = old, new, young: age	S7.1- = (lack of) power/organising

Table 6. Love comedies and tragedies: characteristic semantic categories (rank-ordered)

It is love comedies that are characterised by the most obviously love-related category, 'intimate/sexual relationship'. The love tragedies, by contrast, are characterised by categories far removed from love: 'warfare, etc', 'lack of life, etc', and so on. Closer inspection of the results in the context of the plays reveals many points of interest. For reasons of space, I will just comment on a few. The appearance of 'plants' as highly characteristic of comedies may seem puzzling. In fact, there is a connection with love, as the following extract illustrates (Silvius explains why he loves Phoebe despite the fact that she is a prostitute) (words assigned to the 'plants' semantic category are emboldened):

Silvius So holy and so perfect is my love, And I in such a poverty of grace, That I shall think it a most plenteous **crop** To glean the broken **ears** after the man That the main **harvest reaps** : loose now and then A scattered smile, and that I'll live upon. *As You like It*

More precisely, the connection is a metaphorical one. As Oncins-Martinez (2006) has pointed out, the underlying cognitive metaphor here is SEX IS AGRICULTURE and its sub-mappings include A WOMAN'S BODY IS AGRICULTURAL LAND. Similarly, metaphor accounts for the presence of the semantic category 'sensory: taste', as illustrated in the following example:

Julia Nay, would I were so angered with the same! O hateful hands, to tear such loving words! Injurious wasps, to feed on such **sweet** honey And kill the bees that yield it with your stings! *Two Gentlemen of Verona*

The underlying cognitive metaphor here is LOVE IS FOOD (see Barcelona 1995: 672-673; see also Oncins-Martinez 2006).

12. Conclusions

The main features of my proposed dictionary can be summarised as follows:

- All 'words' will be treated equally (e.g. not just 'hard' words or 'content' words).
- Meanings will not be restricted to semantic or ideational meaning.
- Meanings will be based on usage in context (e.g. not etymology).
- Context will include linguistic co-text (e.g. collocations) and non-linguistic context (e.g. social properties of the speaker).
- Linguistic description will be relative, i.e. it will compare Shakespeare's usage with that of contemporary texts.
- The dictionary will include linguistic profiles of characters and plays.

Perhaps the most important question to raise at this stage is: to what extent is this agenda feasible? In fact, the reason why am from proposing this kind of dictionary now is that until recently it would have been impossible. With developments in both corpora and computational techniques, we are now at a point when it can be realised. To conclude this article, I will briefly list some methodological problems and indicate the extent to which they have been solved:

- There used to be a lack of comparative textual data in electronic form. However, this has been partially solved by, for example, the *Helsinki Corpus of English Texts*, *Corpus of English Dialogues 1560-1760*, and so on.¹³
- Early Modern spelling variation has been perhaps the major stumbling block for historical corpus linguistics, and hitherto the major stumbling block for the proposed dictionary, for the reason that one cannot search on a particular word-spelling and assume

¹³ Regarding Shakespeare's texts themselves, the electronic revolution arrived sometime ago. The best example is probably the Shakespeare Database Project (see <http://www.shkspr.uni-muenster.de/index.php>), although these materials are not publicly available.

that all the relevant words will be retrieved. However, this problem has largely been solved by the Variant Detector (VARD), primarily devised by Dawn Archer (University of Central Lancashire) and Paul Rayson (Lancaster University) (see, for example, Archer et al. 2003, Archer and Rayson 2004).

- Studying abstract grammatical patterns in a corpus requires grammatical annotation. The Lancaster-developed CLAWS part-of-Speech annotation system works fairly well for present-day English (for descriptions of how CLAWS works, see Leech et al. 1994 or Garside 1987). It has been recently adapted at Lancaster for Early Modern English. However, it is not sufficiently accurate for the dictionary and manual correction is required (once this is done, of course, a powerful resource will be created).
- Semantic annotation has received attention from generations of researchers at Lancaster University, including Geoffrey Leech, Jenny Thomas, Roger Garside, Andrew Wilson, Paul Rayson and Dawn Archer. The USAS semantic annotation system has been adapted for Early Modern English, and demonstrated to have value (see, for example, Archer et al. 2003). However, it is not sufficiently accurate for the dictionary, and would require a further round of development. There is also the thorny problem of what 'world view' the system should adopt.
- Social annotation, information about, for example, gender, status, age, has not yet been comprehensively and systematically applied to Shakespeare, but the methodology has been developed and applied to Early Modern English texts (see Archer and Culpeper 2003), and so it would be fairly straightforward to extend this to Shakespeare.
- A final problematic area to note, and one that is philological and not computational, is that the dictionary will need to be based on one particular edition of Shakespeare, and this will involve an evaluation of available editions to arrive at a final choice.¹⁴

References

- Archer, Dawn and Jonathan Culpeper 2003. "Sociopragmatic Annotation: New Directions and Possibilities in Historical Corpus Linguistics." Eds. Andrew Wilson, Paul Rayson and Anthony M. McEnery. *Corpus*

¹⁴ In principle, the dictionary could accommodate multiple editions. However, this would multiply the complexity of the project and be extremely space consuming for the paper dictionary.

- Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Frankfurt: Peter Lang. 37-58
- Archer, Dawn, Tony McEnery, Paul Rayson, Andrew Hardie 2003. "Developing an Automated Semantic Analysis System for Early Modern English." Eds. Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. *Proceedings of the Corpus Linguistics 2003 Conference. UCREL Technical Paper Number 16*. UCREL, Lancaster University. 22-31. <url: [http:// www.comp.lancs.ac.uk/computing/users/paul/public.html](http://www.comp.lancs.ac.uk/computing/users/paul/public.html)>
- Archer, Dawn and Paul Rayson 2004. "Using an Historical Semantic Tagger as a Diagnostic Tool for Variation in Spelling." Presented at 13th ICEHL, University of Vienna. <url: <http://www.comp.lancs.ac.uk/computing/users/paul/public.html>>
- Archer, Dawn, Jonathan Culpeper and Paul Rayson Forthcoming. "Love – 'a Familiar or a Devil'? An Exploration of Key Domains in Shakespeare's Comedies and Tragedies." Ed. Dawn Archer. *What's in a Word-List? Investigating Word Frequency and Keyword Extraction*. London: Ashgate.
- Barcelona Sánchez, Antonio 1995. "Metaphorical Models of Romantic Love in *Romeo and Juliet*." *Journal of Pragmatics* 24: 667-688.
- Biber, Douglas, Susan Conrad and Randy Reppen 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, Douglas and Edward Finegan 1992. "The Linguistic Evolution of Five Written and Speech-based English Genres from the 17th to the 20th Centuries." Eds. Matti Rissanen, Ossi Ihalainen, Terttu Nevalainen and Irma Taavitsainen. *History of Englishes: New Methods and Interpretations in Historical Linguistics*. Berlin: Mouton de Gruyter. 688-704.
- Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad and Edward Finegan 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.
- Blake, Norman F. 2004. *Shakespeare's Non-standard English: A Dictionary of his Informal Language*. Athlone Shakespeare Dictionary series. London: Continuum.
- Boyce, Charles 1996. *Dictionary of Shakespeare*. Ware, Hertfordshire: Wordsworth Editions Ltd.
- Burrows, John F. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and Experiment in Method*. Oxford: Clarendon Press.
- Crystal, David and Ben Crystal 2002. *Shakespeare's Words: A Glossary and Language Companion*. London: Penguin.
- Culpeper, Jonathan 2002. "Computers, Language and Characterisation: An Analysis of Six Characters in *Romeo and Juliet*". Eds. Ulla Melander-Marttala, Carin Östman and Merja Kytö. *Conversation in Life and in Literature: Papers from the ASLA Symposium, Association Suedoise de Linguistique Appliquée (ASLA)*, 15. Universitetstryckeriet: Uppsala. 11-30.

- Culpeper, Jonathan and Merja Kytö 2002. "Lexical Bundles in Early Modern English: A Window into the Speech-related Language of the Past." Eds. Teresa Fanego, Belén Méndez-Naya and Elena Seoane. *Sounds, Words, Texts, Change. Selected Papers from the Eleventh International Conference on English Historical Linguistics (11 ICEHL)*. Amsterdam: John Benjamins. 45-63.
- Enkvist, Nils Erik 1964. "On Defining Style." Eds. Nils Erik Enkvist, John Spencer and Michael Gregory. *Linguistics and Style*, Oxford: Oxford University Press. 1-56.
- Firth, John R. 1957. "A Synopsis of Linguistic Theory 1930-1955." *Studies in Linguistic Analysis*. Oxford: Philological Society. Reprinted in F.R. Palmer ed. 1968. *Selected Papers of J. R. Firth*. Harlow: Longman.
- Foster, John 1908. *A Shakespeare Word-Book: Being a Glossary of Archaic Forms and Varied Usages of Words Employed by Shakespeare*. London: Routledge.
- Garside, Roger 1987. "The CLAWS Word-tagging System". Eds. Roger Garside, Geoffrey Leech and Geoffrey Sampson. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman. 30-41.
- Halliday, M.A.K. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.
- Howard-Hill, Trevor H. 1969-1972. *Oxford Shakespeare Concordances*. Oxford: Oxford University Press.
- Hunston, Susan and Gill Francis 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.
- Kilgarriff, Adam, Pavel Rychly, Pavel SMrz and David Tugwell 2004. "The Sketch Engine." *Proc EURALEX*, Lorient. <url: <http://www.sketchengine.co.uk>>
- Leech, Geoffrey, Roger Garside and Michael Bryant 1994. "CLAWS4: The Tagging of the British National Corpus." *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Kyoto. 622-628. <url: <http://www.comp.lancs.ac.uk/computing/research/ucrel/papers/coling.html>>
- Leech, Geoffrey, Paul Rayson and Andrew Wilson 2001. *Word Frequencies in Written and Spoken English Based on the British National Corpus*. Harrow: Pearson Education.
- Leech, Geoffrey N. 1981. *Semantics*. London: Penguin.
- Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Oncins-Martínez, José Luis 2006. "Notes on the Metaphorical Basis of Sexual Language in Early Modern English." Eds. Juan G. Vázquez-González et al. *The Historical Linguistics-Cognitive Linguistics Interface*. Huelva: University of Huelva Press.
- Onions, Charles T. 1986 (1911). *A Shakespeare Glossary*. (Enlarged and revised by Robert D. Eagleson.) Oxford: Clarendon Press.
- Partington, Alan 1998. *Patterns and Meanings*. Amsterdam: John Benjamins.

- Schiffrin, Deborah 1994 (1987). *Discourse Markers*. Cambridge: Cambridge University Press.
- Schmidt, Alexander 1971 (1902). *Shakespeare Lexicon and Quotation Dictionary* (Vols. 1 and 2). Mineola, NY: Dover Publications.
- Scott, Mike R. 1999. *WordSmith Tools*. Oxford: Oxford University Press. <url: <http://www.liv.ac.uk/~ms2928>>
- Sinclair, John 1987. *Collins COBUILD English Language Dictionary*. London: HarperCollins.
- Sinclair, John 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, John 2004. (edited with Ronald Carter) *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
- Spevack, Marvin 1968-1980. *A Complete and Systematic Concordance to the Works of Shakespeare*. Hildesheim: Georg Olms.
- Stubbs, Michael and Isabel Barth 2003. "Using Recurrent Phrases as Text-type Discriminators: A Quantitative Method and some Findings." *Functions of Language* 10/1: 65-108.
- Wells, Stanley 1998. *A Dictionary of Shakespeare*. Oxford: Oxford Paperbacks.

Author's address:

Department of Linguistics and English Language · Lancaster LA1 4YT · UK
j.culpeper@lancaster.ac.uk