Keyness

Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*[*]

Jonathan Culpeper
Lancaster University

This paper explores keywords, key part-of-speech categories and key semantic categories and their role in text analysis. The first part of the paper addresses a set of issues relating to the definition of keywords and their history, the settings used in deriving keywords, the choice of reference corpora, the different kinds of keyword that emerge in one's results and the dispersion of keywords in one's data. It argues, amongst other things, that keywords are the same as style markers, and that three types of keyword can be identified: interpersonal, textual and ideational. The second part of the paper addresses the question of what precisely is to be gained from analysing key part-of-speech or key semantic domains in addition to keywords. It shows that whilst in general they add little to a keyword analysis, which is in any case methodologically more robust, there are some significant specific benefits. Answers to many of the questions posed in this paper are illustrated by a study of character-talk from Shakespeare's play *Romeo and Juliet*, and in this way this paper also makes a contribution to the fledging field of corpus stylistics.

Keywords: keywords, key parts-of-speech, key semantic domains, style, style markers, *Romeo and Juliet*

1. Introduction

Keyword analysis has a relatively recent history, though it is rapidly gaining steam: for example, Tribble (2000) examines Romantic fiction, Johnson et al. (2003) newspaper political correctness discourse, Baker (2004) gay and lesbian texts, and Xiao and McEnery (2005) spoken and written discourse (see also the papers in Archer 2009). Mike Scott's own publications, as well as those of many other authors, can be found in Mike Scott's comprehensive bibliography available at: http://www.

International Journal of Corpus Linguistics 14:1 (2009), **29–59**. DOI 10.1075/ijcl.14.1.03cul
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

lexically.net/publications/publications.htm. The concept of a keyword, a word that is statistically characteristic of a text or texts, is certainly not new, having a history of at least 50 years, and embryonic keyword analyses were being conducted at least 20 years ago, as Section 2 of this paper will elaborate. What is notably different for more recent times is the advent of computer programs (especially Mike Scott's *WordSmith Tools*, 1996–2008) that perform the required analysis. It is now a relatively easy and rapid task for a researcher to calculate the incidences of each and every single word in the target data as well as a comparative data set, undertake statistical comparisons between incidences of the same words in order to establish significant differences, and finally see the resulting keywords ranked according to degrees of significance of difference. The relative ease and speed of the analytical task is also apparent when one considers other methods designed to reveal styles. Using data comprising three genres (conversation, monologic speech and academic prose), Xiao and McEnery (2005) compared a multi-dimensional analysis of the type conducted in Biber (1988) with a keyword analysis. The results they obtained were "similar" (2005:76) for both methodologies, but keyword analysis was "less demanding" (2005:77), for the reason that multi-dimensional analysis first involves some relatively complex algorithms for the extraction of certain grammatical features from the corpus, and then relatively complex statistical analysis. However, perhaps as a consequence of the ease and rapidity of keyness analysis, some studies perform a keyword analysis in a relatively mechanical way without a critical awareness of what is being revealed or how it is being revealed. The first part of this paper addresses a set of questions relating to keywords, and in so doing aims to raise that awareness. More specifically, it aims (1) to clarify and contextualize the concept of a keyword, (2) to review the role of the statistical settings and reference corpora used in deriving keywords, and (3) to investigate keyword results, proposing that three different kinds of keyword can emerge and also briefly noting the importance of the dispersion of keywords in one's data.

The second, somewhat longer, part of this paper investigates the extension of the notion of keyness to part-of-speech tags and semantic domain tags. Recent developments, most notably Paul Rayson's web-based suite of tools constituting *WMatrix* (Rayson 2005, 2008; see also http://ucrel.lancs.ac.uk/wmatrix), have enabled users to annotate their data sets relatively easily and rapidly for both grammatical and semantic categories, and then to identify which categories are key. Recent studies include: Jones et al. (2004), focussing on key part-of-speech categories in a spoken corpus of English for Academic Purposes, Afida (2007), focussing on semantic domains in business English, and Archer et al. (2009), focussing on semantic domains in Shakespeare's plays (for further references, see http://ucrel.lancs.ac.uk/wmatrix). The particular question this part of the paper addresses is precisely what such studies gain from extending keyness analysis to grammatical or semantic tags. Of course, there are important methodological and theoretical debates regarding the value of annotation. Sinclair (e.g. 2004), for example, is a notable exponent of the view that we should "trust the text" and not sully it with annotations, the analyst's interpretative categories. Instead of condemning such annotation at the outset, this paper takes a more empirical approach by examining the results of such analyses. The three analyses (keyword, key part-of-speech tag and key semantic domain tag) were repeated on the same data. Each and every resulting key item was checked manually, including all items constituting categories, in order to assess whether that item accounted for a textual pattern or style that had or had not been accounted for in the other analyses (and also, in the case of the grammatical and semantic analyses, whether that item was simply a tagging error).

Rayson (2004, 2008) argues for conducting key part-of-speech and semantic domain analyses in addition to keyword analyses, because the former give rise to analytical categories that (1) are fewer than keywords, thus reducing the number of categories a researcher needs to take into account, and (2) group lower frequency words which might not appear as keywords individually and could thus be overlooked. The first point is pitched as a repost to Berber Sardinha's (1999) criticism of keyword analyses, namely, that they deliver more keyword results than is possible for the researcher to analyze. However, another way of partially dealing with this problem is simply to change the keyword settings so that fewer keywords are derived (see Section 3). The second point raises a more fundamental issue. However, it is not clear how much of an issue this is: how often do groups of lower frequency words which would not appear independently as keywords emerge? Rayson (2008) does not display all the items that constitute the part-of-speech or semantic categories, so one cannot tell the extent to which they overlap with keywords. The second part of the paper aims at such detailed consideration and comparison of what each analysis reveals, and also suggests why differences emerge.

This paper builds on the keyword analysis of Shakespeare's *Romeo and Juliet* reported in Culpeper (2002). The same data will be used for all three keyness analyses, that is, the speech of the six characters who speak the most in Shakespeare's *Romeo and Juliet*, their total speech varying from 5,031 to 1,293 words. These data represent a good test-bed for the analyses for two reasons. One is that the text for each character is highly likely to constitute a different, and sometimes radically different, kind of style, if we accept it is the speech of each character that partly determines the different characters we perceive. Conducting the analyses on a range of styles obviously makes for a more comprehensive test, more likely to expose various strengths and weaknesses (Rayson 2008, for example, uses only one genre, political manifestos). The other is that the small datasets mean that it is possible to scrutinise carefully and manually all the results from the analyses. It is also the case that in the process of conducting analyses and accounting for the results this paper makes a contribution, especially in Sections 4 to 7, to the fledging field of corpus stylistics, a field that focuses on the application of techniques from corpus linguistics to literary texts and the interpretation of the results (see, for example, Wynne 2006). Shakespeare's plays have the advantage that they are relatively well known throughout the world, and thus many readers will be able to relate at least to some degree to the characters, particularly the major ones, discussed. However, it has the disadvantage that it is historical text, though the text used here is the modernised Craig (1914) Oxford edition. I will very briefly point out issues pertaining to the historical nature of the text during the paper.

2. What are keywords?[1]

Needless to say, the term 'keyword' is not to be confused with lexical items that are 'key' because they are of particular social, cultural or political significance (see for example, Williams 1976). It is simply a term for statistically significant lexical items. Studies in the area of stylometry have long known of statistically significant items, though perhaps the first to use the term keyword ('mots-clés') for this particular concept was Pierre Guiraud (1954). Guiraud (1954: 64–66) contrasts 'mots-clés' (based on relative frequency) with 'mots-thèmes' (based on absolute frequency):

> Toute différente est la notion de *mots-clés*, qui ne sont plus considérés dans leur fréquence absolue, mais dans leur fréquence relative : ce sont les mots dont la fréquence s'écarte de la normale. [Wholly different is the notion of *mots-clés* (keyword), which are not considered in terms of their absolute frequency, but their relative frequency; these are the words whose frequency diverges from the normal.]

Simply being statistically significant is not in itself the important point of interest. That lies in the link between keywords and style. Although he does not use the label keywords, this link is clearly articulated by Nils Erik Enkvist (e.g. 1964). In the following quotations, Enkvist defines style in terms of 'frequencies', 'probabilities' and 'norms', and goes on to define 'style marker':

> Style is concerned with frequencies of linguistic items in a given context, and thus with *contextual* probabilities. To measure the style of a passage, the frequencies of its linguistic items of different levels must be compared with the corresponding features in another text or corpus which is regarded as a norm and which has a definite relationship with this passage. For the stylistic analysis of one of Pope's poems, for instance, norms with varying contextual relationships include English eighteenth-century poetry, the corpus of Pope's work, all poems written in English in rhymed pentameter couplets, or, for greater contrast as well as comparison, the

> poetry of Wordsworth. Contextually distant norms would be, e.g., Gray's *Anatomy* or the London Telephone Directory of 1960. (1964:29)

> We may […] define style markers as those linguistic items that only appear, or are most or least frequent in, one group of contexts. In other words, style markers are contextually bound linguistic elements. Elements that are not style markers are stylistically neutral. This may be rephrased: style markers are mutually exclusive with other items which only appear in different contexts, or with zero; or have frequencies markedly different from those of such items.
>
> In the light of this, some otherwise meaningless repetitions of linguistic items acquire meaning as style markers. For instance, the swearing and cursing of a soldier introduces a stream of stylistically significant items — 'style reminders' — into statements that would otherwise remain neutral. (1964:34–5)

Style markers as words whose frequencies differ significantly from their frequencies in a norm are precisely what keywords are. Repetition is the notion underlying both style markers and hence keywords, but not all repetition, only repetition that statistically deviates from the pattern formed by that item in another context.

Using the notion of style markers or keywords to reveal the textual patterns or 'styles' in particular data is not new either. For example, more than 20 years ago, Burrows (1987) examined the vocabulary of characters in Jane Austen's novels in order to identify distinctive styles, using a variety of statistical measures, including cross-tabulation and chi-square — the very statistics that underlie most keyword analyses (see 1987: Chapter 2). Burrows did deploy a computer, but even so he was reduced to examining a mere three words, *we*, *our* and *us*, retrieving incidences for every character and then calculating significant differences. It is, however, in the context of corpus linguistics that the notion of keywords and the practice of keyword analysis has been developed and popularised, most notably by Mike Scott through the *KeyWords* facility of his program *WordSmith Tools*, a program designed for the computational analysis of corpora.[2] This program conducts a statistical comparison between the words of a corpus (or wordlist) and a bigger comparative or reference corpus, in order to identify words that are unusually frequent or unusually infrequent or, in other words, keywords. According to Scott (2008:144; the punctuation is not original):

> To compute the "key-ness" of an item, the program therefore computes its frequency in the small wordlist, the number of running words in the small wordlist, its frequency in the reference corpus, the number of running words in the reference corpus and cross-tabulates these. Statistical tests include: the classic chi-square test of significance with Yates correction for a 2 X 2 table; Ted Dunning's Log Likelihood test, which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus. A word

> will get into the listing here if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger wordlist.

'Keyness' is a matter of being statistically unusual relative to some norm. The statistical operations involved here, a cross tabulation and a chi-square or log likelihood significance test, are basic and commonly used in corpus linguistics.

3. What decisions need to be made in performing a keyword analysis?

In any keyword analysis, the choice of data for comparison (the reference corpus) is an issue. There is no magic formula for making this decision. Scott and Tribble (2006:58) suggest that it "should be an appropriate sample of the language which the text we are studying (the "node-text") is written in". As for what constitutes an "appropriate sample", they go on to say that it "usually means a large one, preferably many thousands of words long and possibly much more" (2006:58). Precisely what counts as large enough is still a matter of debate. Xiao and McEnery (2005:70) compared two reference corpora, the 100-million-word British National Corpus and the one-million-word Freiburg-LOB Corpus, and achieved almost identical keyword lists, thus concluding that "the size of the reference corpus is not very important in making a keyword list". Similarly, Scott and Tribble (2006:64), experimenting with various reference corpora for a comparison with the play *Romeo and Juliet*, concluded that "while the choice of reference corpus is important, above a certain size, the procedure throws up a robust core of KWs whichever reference corpus used". Even so, a set of data which has no relationship whatsoever with the data to be examined is unlikely to reveal interesting results regarding the stylistic characteristics of that data (cf. Enkvist's comparison of a poem by Pope with a telephone directory). What if one simply selects a huge multi-genre corpus, such as the British National Corpus, as indeed other studies have done (e.g. Tribble 2000; Scott 2000; Johnson et al. 2003)? In this case we can readily hypothesize that some genres within that corpus have a relatively close relationship with the data to be examined, but other genres have a relatively distant relationship. These relationships will influence the keywords revealed.

Let us consider a study that used such a multi-genre corpus, Johnson et al. (2003). Here, the data to be examined consisted of newspaper articles which contained political correctness expressions (e.g. *political correctness*, *politically correct*, *politically incorrect*), the research interest being to discover what characterised the discourse in which those expressions appeared. These data were compared with a word-list based on the entire BNC multi-genre set of written texts (90.7 million words). Amongst the most key keywords were *is*, *has*, *who* and *says*. These were frequent items in the political correctness corpus and evenly dispersed. However, upon close analysis, no connection with political correctness matters could be discerned, despite the fact that the target newspaper data had been selected because it was characterised by political correctness discourse. Two of the items, *who* and *says*, were found by Biber et al. (1999: 375, 610) to be outstandingly frequent in newspaper language generally. Thus the problem with the resulting list of keywords is that some reflected newspaper discourse in general as opposed to political correctness discourse in particular. What this suggests then, is that the choice of the reference corpus will affect whether you acquire keyword results that are all relevant to the particular aspect of the text(s) you are researching. The 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. Thus, for the above study, a comparison with a corpus of newspaper texts (excluding political correctness-related texts) should have provided results specific to political correctness discourse, as features of newspaper discourse in general would most likely be common to both the target and reference corpus, and therefore not be identified as key. The issue is how specific you want *all* your keyword results to be.

In my *Romeo and Juliet* analyses in the sections below, the comparative reference corpus was the speech of the six characters minus the one being investigated (e.g. Romeo's speech was compared with the speech of the other five characters). This contrasts with Scott and Tribble's (2006: Chapter 4) analysis of the play, which used all of Shakespeare's plays as a reference corpus, and thus derives a somewhat different set of keyword results. The analyses here produce key items that reflect the distinctive styles of each character compared with the other characters in the same play, rather than — if one had compared them with all Shakespeare's plays — stylistic features relating to differences of genre (e.g. the fact that the play is a tragedy rather than a history or comedy) or aspects of the fictional world (e.g. the aristocratic Italian settings rather than the royal court of the English history plays).

A *Keywords* program (within *WordSmith Tools*, for example) usually allows the user to set various parameters. One such parameter is a minimum frequency cut-off point. The point of this parameter is to exclude words that will be identified as unusual simply because they happen not to have occurred or to have occurred very infrequently in the dataset or reference corpus. Proper nouns, for example, are often amongst these one-off occurrences. This is not to say that such phenomena — which are referred to as 'hapax legomena' — are uninteresting (see, for example, Hoover 1999: Chapter 4). The problem is that in a list of keyword results, mixing frequent items with the very infrequent, often means mixing generalised phenomena with extremely localised, which has the result of making an account of a keyword list problematic. Setting the minimum frequency cut-off at 10 is popular, but this could lead to very few resulting keywords if the size of your data set is small. In the *Romeo and Juliet* study referred to in this paper, the minimum frequency for a word to be considered key is set at five, because of the relatively small data set.

Another parameter is the test for statistical significance. The point of the significance test is that it calculates the significance of the unusualness of the keyword. Almost every word of a corpus will have some difference in frequency from what one might expect on the basis of the reference corpus. The significance test enables one to assess the strength of those differences. In the study discussed below, I selected the log-likelihood test for significance, but I repeated the analysis with the chi-square test. The same results were revealed with only minor and occasional differences in the ranking of keywords, differences which had no effect on the overall picture revealed by the keywords. I set the probability value at smaller than or equal to 0.01. Thus, words whose differences were considered to have a 1% chance or less of being a fluke would be included as keywords; words with more of a chance of being a fluke would be excluded. Rayson (2003), evaluating various statistical tests for data involving low frequencies, different corpus sizes and so on, favours the log-likelihood test 'in general' and, moreover, a 0.01% significance level "if a statistically significant result is required for a particular item" (2003:155). However, Scott (2008:145–6) points out that "with keywords where the notion of risk is less important than that of selectivity, you may wish to set a comparatively low p value threshold such as 0.000001 (1 in a million) (1E-6 in scientific notation) so as to obtain fewer keywords". Manipulating the *p* value is a useful way of controlling the quantity of keywords derived, and thus the number of keywords a researcher must interpret.

In sum, and rather like Baker (2004), my approach to settings is to derive them by testing various possibilities and, in most cases, choosing a combination that results in: (1) a sufficient number to meet one's research goals, (2) a not overwhelming number of words to analyse, (3) an adequate dispersion of at least some keyword instances, and (4) any one-off or extremely rare word types being minimised. Of course, whilst the previous sentence identifies important factors, there is little clarity regarding what counts as "sufficient", "not overwhelming", "adequate" or "minimised". It may be possible for future research to produce more precise guidelines, though settings cannot be reduced to a simple mathematical formula for the reason that different research purposes and contexts have different requirements.

4. What kinds of keyword result from an analysis?

Let us consider the keywords for the six characters in Shakespeare's *Romeo and Juliet* who speak the most (the settings used for this analysis were given in the previous section). For each character, Table 1 presents positive keywords that appear because they are unusually frequent and negative keywords that appear because they are unusually infrequent. The fact that there are fewer negative keywords compared with positive keywords is not surprising: it is easier to do more than the

Table 1. Keywords for six characters in *Romeo and Juliet* (in descending order of keyness, with frequency of occurrence given in brackets)[3]

	Romeo	**Juliet**	**Capulet**	**Nurse**	**Mercutio**	**Friar L.**
Positive keywords	Beauty (10)	If (31)	Go (24)	Day (22)	A (85)	Thy (51)
	Love (46)	Be (59)	Wife (10)	He's (9)	Hare (5)	From (23)
	Blessed (5)	Or (25)	You (49)	You (55)	Very (11)	Thyself (5)
	Eyes (14)	I (138)	Ha (5)	Quoth (5)	Of (57)	Her (30)
	More (26)	Sweet (16)	Thank (5)	God (12)	The (85)	Mantua (6)
	Mine (14)	My (92)	Her (29)	Woeful (6)	He (20)	Part (7)
	Dear (13)	News (9)	T (5)	Warrant (7)	O'er (5)	Heaven (10)
	Rich (7)	Thou (71)	Thursday (7)	Madam (10)	An (14)	Forth (5)
	Me (73)	Night (27)	Child (7)	Lord (11)	Eye (5)	Alone (6)
	Yonder (5)	Would (20)	Welcome (5)	Lady (16)	Us (7)	Time (10)
	Farewell (11)	Yet (18)	We (15)	It (39)		Married (7)
	Sick (6)	That (82)	Tis (11)	Hie (5)		Thou (46)
	Lips (9)	Nurse (20)	Haste (6)	Your (21)		In (51)
	Stars (5)	Name (11)	Gentlemen (5)	Faith (7)		Then (18)
	Fair (15)	Words (5)	Our (13)	She (21)		Letter (5)
	Hand (11)	Tybalt's (6)	Make (10)	Ay (90)		
	Thine (7)	Send (7)	Now (15)	Said (6)		
	Banished (9)	Husband (7)	Well (13)	About (5)		
	Goose (5)	Swear (5)	Daughter (5)	Sir (13)		
	That (84)	Where (16)		Ever (5)		
		Again (10)		Marry (7)		
				A (61)		
				Ah (6)		
				O (26)		
				Well (13)		
				Fall (5)		
				Mother (5)		
Negative keywords	He (11)	The (84)	The (37)	Thou (11)	What (5)	Have (5)
	Romeo (5)	You (27)	That (13)		I (31)	A (33)
	You (14)	Her (5)	Thou (7)		My (13)	You (16)
						I (32)

norm established in a reference corpus than do less than that norm, particularly when the reference corpus is small.

Scott comments in a number of publications (e.g. 2000, 2008; Scott & Tribble 2006) that keywords tend to be of two main types, with a possible third, with which we will begin. Firstly, there are proper nouns. Scott's suggestion is that these are of little importance: "a text about racing could wrongly identify as key, names of horses which are quite incidental to the story" (2008:143). In fact, in fictional texts, they may be of some interest, as they relate to key aspects of the fictional world. However, in Table 1 they are very few — merely three — and two are highly localized (as discussed in the following section). Interestingly, the proper noun *Romeo* is a *negative* keyword for Romeo; proper nouns do not appear as negative keywords for any other character. This reflects the fact that *Romeo* is a frequent term of address or reference for other characters, but not used frequently in self-reference by Romeo himself (only 5 instances). This is some evidence that Romeo is the fulcrum of the play.

Secondly, there are keywords that relate to the text's 'aboutness' (a term used in, for example, Phillips 1989) or content. Scott (2000:155) relates aboutness to Halliday's (e.g. 1994) ideational metafunction, and also suggests that they "are key words that human beings would recognise" (2008:143) or would be "likely to predict" (2000:160). Romeo's most key keywords illustrate this well: surely most people would guess that Romeo's talk was about *beauty* and *love*! Aboutness keywords are also involved in the construction of Capulet, although they also have an important grammatical dimension — the imperative mood. His keywords (e.g. *go*) help characterise not only his social position as 'director' of the household but also his dramatic one as a character set up for a tragic fall — the person he fails to direct is his own daughter.

Thirdly, there are "indicators more of style than of 'aboutness'" (2008:143), and Scott cites such examples as *because*, *shall* and *already*. Style seems to be a cover-term for items not obviously indicating aboutness (Scott & Tribble 2006:60). Generally, it appears to be the case that aboutness keywords relate to 'open class' words, whilst stylistic keywords relate to 'closed class' words.[4] Juliet's most key keywords illustrate this well: *if*, *be*, *or*, and *I* are all frequently occurring items that most people would be unlikely to predict. Here are some examples (keywords are underlined):

> If he be married, / Our grave is like to be our wedding-bed (I.v.) [at her first sighting of him, whether Romeo is married]
>
> If they do see thee, they will murder thee (II.ii.) [whether Romeo will be spotted during a covert visit]
>
> But if thou meanest not well (II.ii.) [whether his intentions are honourable and his love will lead to marriage]

The keyword *if* can be accounted for by the fact that Juliet is in a state of anxiety for much of the play. But it is not just this keyword. *If* works together with other keywords — *be* (almost always subjunctive), *or*, *yet, would* — which are also more grammatical in nature to create a particular grammatical style that can be related to the anxieties we perceive in Juliet. Mercutio's speech, even more than Juliet, is characterised by stylistic keywords, most being grammatical, including *a*, *very*, *of*, *the* and *an*. Unlike Juliet, Mercutio's keywords do not specify logical semantic relations, but are part of a highly rhetorical style which deploys lists of noun phrases (e.g. *the very / butcher of a silk button, a duellist, a duellist; a gentleman of the very first / house, of the first and the second cause. / Ah, the immortal passado! the punto reverso! / the hay!* II.iv.). This can be related to an impression one gets of Mercutio as all style and no substance.

A problem with the distinction between aboutness and stylistic keywords is that, whilst useful, it can lead one to assuming a simple "dualist" view of style, whereby choices of content are separable from stylistic choices — style is merely decoration (see Leech & Short 2007: Chapter 1). A functional, Hallidayan approach, for example, would view all choices, including grammatical choices, as meaningful and stylistic.[5] Furthermore, the Nurse's keywords are not easily categorisable as aboutness or stylistic keywords. Discourse markers and interjections such as *warrant*, *faith*, *marry*, *ah*, *o* and *well*, and vocatives such as *god*, *madam*, *lord*, *lady* and *sir* are the Cinderellas of language, as they are considered by some linguists not to be part of the grammar or the lexicon. Discourse markers, and to some extent vocatives, have little semantic content, but rather pragmatic import, and they tend to be peripheral to the syntax. Given neither the term aboutness nor the term stylistic does justice to the Nurse's keywords and the problem to do with the view of style suggested by the term aboutness contrasting with stylistic, I suggest that we adopt a three-way categorisation of keywords, broadly following Halliday (e.g. 1973, 1978, 1994). This consists of 'ideational keywords' (encompassing Romeo in particular), 'textual keywords' (encompassing Mercutio) and 'interpersonal keywords' (encompassing the Nurse). Needless to say, these are not discrete categories; they are designed to capture a functional emphasis.

5. Are all keywords general features of the data in focus?

The objective with a keyword analysis is to make a claim that a certain set of words is key to a certain set of data relative to a comparative reference corpus. The problem is that it is easy to retrieve keywords that are key, but not actually general features of the data one is examining. This can lead to some highly misleading characterisations of particular discourses or genres. I have already mentioned one way of minimising highly local or idiosyncratic keywords, and that is to institute a frequency cut-off point. Minimally, one can adopt the good practice of giving raw frequencies of particular items in keyword lists, as I did in Table 1. Clearly, keywords with lower frequencies are more likely to be suspect, that is, clustered in a certain part of the data. But one can also look at dispersion more directly. To illustrate, I will focus on characterisation and Romeo's keywords. An important factor — though not necessarily a decisive one — in determining whether keywords relate to a particular character or not is whether they are localised or well-dispersed throughout the play.

Regarding Romeo, note that it is not until Act I scene v that Romeo notices Juliet; prior to this, Rosaline is the subject of his infatuation. This has some implications for the way the keyword instances are dispersed across the play. *WordSmith Tools* usefully allows one to generate a dispersion plot, as I did for Romeo in Figure 1 (the keywords in bold at the bottom are negative keywords; each file contains a different scene, and the files/scenes involved in Romeo's keywords are indicated at the top of the figure).

Love is dispersed widely, appearing in every scene that Romeo does, except two: clearly this is a consistent, general feature of his characterisation. However, one can see something of a concentration in two scenes: in Act I scene i, Romeo extols his love for Rosaline to his cousin Benvolio; in Act II scene ii, he extols his love for Juliet, who appears on the balcony. *Beauty* is used of both women, but *blessed* only of Juliet, who is metaphorically deified as the object of his love. Figure 1 also shows that two keywords are highly localised. Romeo's keyword *banished* only occurs

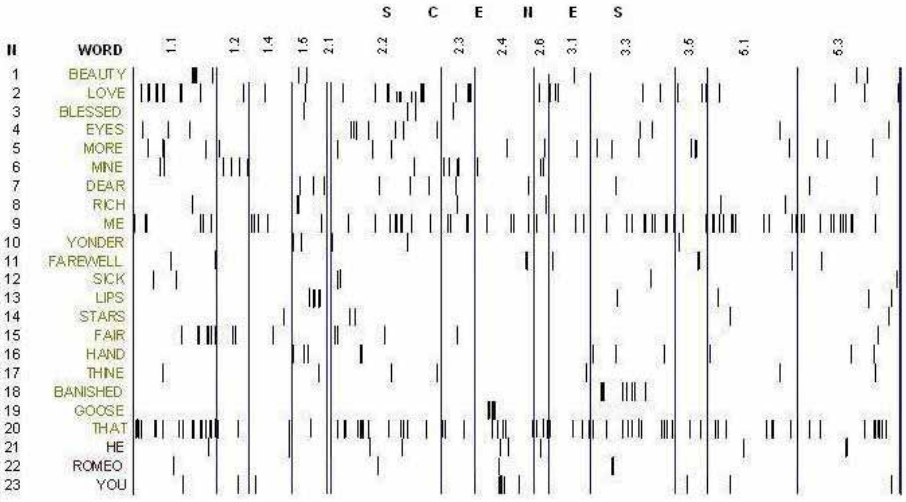

Figure 1. The dispersion of Romeo's keywords

in Act III scene iii: it is a localised reaction to the circumstances he finds himself in and not a general feature of his character. *Goose* only occurs in Act II scene iv, where Romeo word plays with Mercutio about a "wild-goose chase". In contrast to Romeo's keywords, Juliet's keywords are fairly evenly dispersed throughout the play: anxiety is a general characteristic of Juliet. Even *yet* with only 18 occurrences occurs in 6 scenes, with just a slight preponderance in Act II scene ii.

One area not considered in this paper is relationships between keywords. Scott (e.g. 1997) has developed, for example, the notion of 'key-keywords' (not simply words that are more key than other keywords, but words that are keywords in a number of different files, i.e. they are generally key across the body of data), and 'associates' (keywords that have a statistical association with other keywords). It was not possible to pursue these notions in this data, on account of the small number of words in each character file. Had it been possible, I would have been able to discuss matters related to dispersion in a more precise and systematic way, rather than simply "eye-balling" dispersion plots. Readers can find a discussion of these issues, as applied to the entire *Romeo and Juliet* play, in Scott and Tribble (2006:66–69, Chapter 5).

6. What is to be gained from analysing key parts-of-speech in addition to keywords?

It is already clear from Section 4 that keyword analysis offers some insights into grammatical style, but how does this compare with treating grammar more explicitly? In order to incorporate grammar (and semantics for that matter) explicitly into a keyness analysis, the data needs to be annotated (i.e. each word needs interpretative grammatical information). Automated, or even semi-automated, tagging is doomed to failure if the spellings of the text are variable. This is a pertinent issue for historical texts, as spelling standardisation was not largely complete until towards the end of the 17th century. My solution was to use the program *Variant Detector* (VARD) (see Archer et al. 2003; Archer & Rayson 2004; Rayson et al. 2005). This program regularises variation by matching variants to "normalised" equivalents using a search and replace script, as well as contextual information to tackle ambiguities and an additional lexicon to treat word forms that are specific to or have undergone semantic change since the Early Modern period.

For part-of-speech annotation, I used the CLAWS (Constituent Likelihood Automatic Word-tagging System) software at Lancaster University (for descriptions of how CLAWS works, see Leech et al. 1994 or Garside 1987).[6] Once this was done, I could analyse the keyness of grammatical tags, and then compare my results with those of my keyword analyses in the previous sections, and address the question in the heading of this section. In practice, I did not have to run my text through the grammatical tagging program and the semantic tagging program, described in Section 7, separately, and then feed the annotated text into *WordSmith Tools*. Instead, I used the much more convenient option of *Wmatrix*. Texts uploaded into *WMatrix* are automatically run through two programs which apply grammatical and semantic annotation, and then within *WMatrix* one can retrieve keyness lists. Here and in the next section, I will focus on Romeo, Mercutio and the Nurse. It may be remembered that the keyword results for these characters offer a full range of keyword types, ideational (Romeo), textual (Mercutio) and interpersonal (Nurse).

As a preliminary to the upcoming analyses, I consider a key dimension of variation in the lexicon, as this is likely to play a role in the nature of the results. The lexicon is understood as an inventory of units varying along a continuum running from the more lexical to the more grammatical (see, for example, Brinton & Traugott 2005). On the more lexical side we have parts-of-speech such as nouns, lexical verbs, and adjectives, whilst on the more grammatical side we have parts-of-speech such as determiners, prepositions, pronouns, conjunctions and auxiliary verbs. The category of adverbs represents a varied category, some items of which (e.g. *very*) are more grammatical, whilst others are more lexical (e.g. *certainly*). The distinction between more lexical and more grammatical units could be referred to as a distinction between more open and more closed class units, a distinction referring specifically to the fact that the class of more grammatical items does not readily accept new members. However, there are other important differences. More lexical items tend to:

- be more "contentful", whereas more grammatical items are more "functional" (for example, a definition of *table* would elaborate on the nature of a concrete object, whereas a definition of *of* would elaborate on its grammatical function in the text);
- have a relatively wide range of types (for example, the category of nouns has a huge range of types, whereas the category of English determiners is relatively restricted); and
- have a relatively low frequency of tokens for any particular type (grammatical items dominate the most frequent items in English).

These characteristics will play a role in the keyness analyses, particularly the part-of-speech analyses, as we shall see.

Table 2 displays the grammatical categories that are key in Romeo's speech (the statistical criteria for all keyness tables in this section and the next are the same as for the *Romeo and Juliet* keyword results in the previous sections, that is, a log likelihood value of 6.63 or higher, which is equivalent to $p < 0.01$, and a raw frequency value of five or more).

Table 2. Romeo's parts-of-speech rank-ordered for positive keyness (i.e. relatively unusual over-use) (keywords, as listed in Table 1, are emboldened)

Grammatical category, including the tag code and frequency	**Items within the category (and their raw frequencies) up to a maximum of ten types if they are available (excluding clear tagging errors in square brackets)**
Nominal possessive personal pronoun (e.g. *mine, yours*) (PPGE)[7] (17)	***mine*** (8), *hers* (4), ***thine*** (3), [*his* (1)], *yours* (1)
Comparative after-determiner (e.g. *more, less, fewer*) (DAR) (16)	***more*** (15), *less* (1)
1st person sing. objective personal pronoun (i.e. *me*) (PPI01) (73)	***me*** (73)
General adjective (JJ) (328)	***fair*** (14), *good* (10),[8] ***dear*** (10),[9] *sweet* (8), ***rich*** (7), *dead* (6), *holy* (5), *true* (5), *heavy* (5), ***blessed*** (4)
1st person sing. subjective personal pronoun (i.e. *I*) (PPIS1) (144)	*I* (144)
Than (as conjunction) (CSN) (16)	*than* (16)

Note that all the key grammatical categories in Table 2 are dominated by a single item, except the category General adjective (JJ). This is not surprising because, as we noted above, such categories have a relatively reduced range of types but high frequencies of tokens. General adjective (JJ), in contrast, is a more lexical category, containing a wider and more even range of items, and also a more contentful category. Is it the case that the more grammatical categories are more likely to include items which have already appeared in the keywords analysis? One might hypothesize that this will be the case, because such categories are dominated by high-frequency items that are selected in a keyword analysis. Of the most frequent words for each of the six categories four (i.e. *mine*, *more*, *me*, *fair*) are indeed also keywords. However, this *includes* the category General adjective (JJ) characterised by a relatively wide range of types and contentful items, although one might add that the status of this entire category is called into question by the fact that the bulk of instances of *good* and *dear* appear in vocative expressions. Conversely, two much less wide-ranging (each comprised of one word) and less contentful categories, 1st person singular subjective personal pronoun (i.e. *I*) (PPIS1) and *than* (as conjunction) (CSN), do not include members which were keywords at the significance level used in this paper (i.e. $p < 0.01$).[10] So, it is not the case that the dominance in grammatical categories of a limited number of high-frequency items guarantees that they will appear in a keyword analysis, though there is a tendency.

Regarding the category General adjective (JJ), note that as there is not a preponderance of adjectives amongst Romeo's keywords in Table 1, this feature of his style could easily have been overlooked, if one had relied solely on the list of keywords. As an example of Romeo's distinctively frequent use of adjectives, consider one of his most famous lines (adjectives are underlined): *If I profane with my unworthiest hand This holy shrine, the gentle sin is this; My lips, two blushing pilgrims, ready stand To smooth that rough touch with a tender kiss* (I.v.).

The fact that the category General adjective (JJ) is key is consistent with the earlier characterisation of Romeo being a character generally characterised by ideational keywords. A total of eight of his keywords listed in Table 1 also occur amidst the grammatical categories in Table 2, with four in the more lexical category General adjective (JJ). But this is only half of the keywords. This reminds us that the description of Romeo's keywords being characterised by ideational keywords is a generalisation: his keywords also contain clues to his grammar. The items *mine* and *me*, for example, were already apparent in Romeo's keyword list.

Table 3 displays the grammatical categories that are key in the Nurse's speech.

Table 3. The nurse's parts-of-speech rank-ordered for positive keyness (i.e. relatively unusual over-use) (keywords, as listed in Table 1, are emboldened)

Grammatical category, including the tag code and frequency	**Items within the category (and their raw frequencies) up to a maximum of ten types if they are available (excluding clear tagging errors in square brackets)**
3rd person sing. subjective personal pronoun (*he, she*) (PPHS1) (46)	***he*** (24) ***she*** (22)
Singular letter of the alphabet (e.g. *A,b*) (ZZ1)[11] (16)	[***o*** (9)], [*a* (3)], [*I* (3)], *r* (1)
Temporal noun, singular (e.g. *day, week, year*) (NNT1) (31)	***day*** (18), *night* (5), *year* (2), [*well-a-day* (1)], *second* (1), *lammas-eve* (1), *hour* (1), *time* (1), *afternoon* (1)
3rd person sing. neuter personal pronoun (*it*) (PPH1) (41)	***It*** (40), *'t* (1)
Interjection (e.g. *oh, yes, um*) (UH) (42)	***o*** (16), ***ah*** (6), ***ay*** (5), *nay* (4), *alas* (3), *no* (2), *amen* (1), *ho* (1), *yes* (1), [*the-no* (1)], *fie* (1), *farewell* (1)

Here, we see a general pattern whereby the item dominating the category also appears in the keyword analysis. This is most starkly the case for the more grammatical categories, third person singular subjective personal pronouns and the third person singular neuter pronoun, for which all instances are accounted for amongst the keywords except for one. Conversely, and as with the category General adjective (JJ) discussed above, the category Temporal noun seems to reveal a grammatical feature of style that was not apparent from the keyword analysis. However, note the dominance of *day*, a keyword which is highly localised to where the Nurse discovers Juliet is dead (*o woeful day*, repeats the Nurse). If this item were excluded, this grammatical category might not turn out to be key. Discourse markers, which include the subcategory of interjections, were clearly established as a feature of the Nurse's speech during the keyword analysis. The top three items in this category were also keywords. Interestingly, in this category we see items that were not keywords (i.e. *nay*, *alas*, *no*, *amen*, *ho*, *yes*, *fie* and *farewell*), and, conversely, we do not see discourse markers that were keywords (i.e. *well*, *marry*, *hie*, *God*, *warrant*). The fact that some items were not also keywords is presumably a consequence of the fact that as individual items they are more evenly distributed among the characters. The fact that some of the keyword interjections are not also represented in the grammatical category above suggests a failure of the tagger (and most probably its lexicon) to identify items like *marry*, *hie* and *warrant* as discourse markers.

Table 4 displays the grammatical categories that are key in Mercutio's speech.

Table 4. Mercutio's parts-of-speech rank-ordered for positive keyness (i.e. relatively unusual over-use) (keywords, as listed in Table 1, are emboldened)

Grammatical category, including the tag code and frequency	**Items within the category (and their raw frequencies) up to a maximum of ten types if they are available (excluding clear tagging errors in square brackets)**
Singular article (e.g. *a, an, every*) (AT1) (96)	***a*** (82), ***an*** (14)
Plural common noun (e.g. *books, girls*) (NN2) (99)	*houses* (4), *dreams* (4), *eyes* (3), *wits* (3), *ears* (3), *maids* (2), *wings* (2), *cats* (2), *bons* (2), *minstrels* (2)
Of (as preposition) (I0) (57)	***of*** (57)
Singular cardinal number (i.e. *one*) (MC1) (10)	*one* (9), [*I* (1)]
Article (e.g. *the, no*) (91) (AT)	***the*** (84), *no* (7)

As above, more part-of-speech categories are dominated by particular items which are also keywords. It is of no surprise to see the definite and indefinite articles and the *of* preposition as key part-of-speech categories. Such keywords had led us to conclude that Mercutio had a nominal style, on the basis that nouns tend to follow such items. This conclusion is supported by the appearance of plural common nouns in Table 4. The grammatical category of noun is much less grammatical in character, and this corresponds with the fact that it is less dominated by keywords — indeed no keywords appear in this category. The fact that the category of common nouns refers specifically to plural nouns is something not predicted by my keyword analysis, but is consistent with the rhetorical generalisations that Mercutio has a taste for.

Many of the results of the keyness analysis of part-of-speech categories are already apparent in the keyword analysis. This seems to be because the part-of-speech categories are often dominated by one or two items, particularly if the part-of-speech categories are more grammatical in character. This, in turn, is because the more grammatical categories are dominated by a restricted set of frequently occurring word-form items that tend to crop up as keywords anyway. The part-of-speech analysis is most revealing in the case of more lexical part-of-speech categories, as illustrated by general adjectives for Romeo and plural common nouns for Mercutio. Here it offers evidence of aspects of a character's style, aspects which had not been apparent in the keyword analyses. The results for the Nurse, and specifically the discourse markers, remind us that automated annotation systems need further development. This is not surprising, as interpersonal items are not well accommodated in grammatical descriptions.

7. What is to be gained from analysing key semantic domains in addition to keywords?

Perhaps a keyword analysis is misleading because the semantic similarities between words are not explicitly taken into account in the analysis? One can manually analyse the semantic similarities of the keyword results, as indeed I did, but those results were not selected on the basis of semantic similarities. Semantic annotation is closely related to 'content analysis', which is "concerned with the statistical analysis of primarily the semantic features of texts" (Wilson & Rayson 1993:2). Analysing literary texts for meaning or content by adding annotations is no alien activity for the literary scholar. But note that content analysis involves statistical analysis, which suggests something altogether more systematic and rigorous. The final processing stages that *WMatrix* conducts on one's texts deploy UCREL's (University Centre for Computer Corpus Research on Language, based at Lancaster University) Semantic Analysis System (USAS), an annotation programme designed for automatic dictionary-based content analysis (http://ucrel.lancs.ac.uk/usas/) (see Rayson et al. 2004). The input to USAS is part-of-speech tagged text as produced by CLAWS, then the program SEMTAG assigns semantic tags (or tags in the case of ambiguities) to each lexical item or multiword unit by matching the words of the data with lexicons (for details see Wilson & Rayson 1993, 1996). Piao et al. (2004) evaluated the coverage of these lexicons by testing it on the British National Corpus. They claim that 99.39% of the BNC spoken data and 97.6% of the written data is covered, though acknowledge that the lexis that is uncovered could always prove critical for corpus analysis. SEMTAG is claimed to achieve an accuracy rate of 91% (Rayson et al. 2004). Still, it must be acknowledged that the reliability of the tagging is an issue. Originally, the tagset used by SEMTAG was based on Tom McArthur's *Longman Lexicon of Contemporary English* (1981), because it offered what seemed the most appropriate thesaurus-type classification of word senses, but the tagset has received significant revisions over the years. The classification has a hierarchical structure with 21 major semantic fields; each top-level category has a letter associated with it; subdivisions are indicated by numerals, and sub-subdivisions by a point and further numerals, and so on (the tagset can be found at: http://www.comp.lancs.ac.uk/computing/research/ucrel/usas/).

There are, of course, problems for historical data. The lexicon changes over time. Piao et al. (2004) report 94.4% coverage of the lexis in the Lancaster Corpus of Seventeenth-Century Newsbooks. However, this is *without* the intervention of the regularising program VARD and its historical lexicon, and so the prospects for the Shakespearean data here may be better. A more fundamental problem relates to the nature of the classification. The classification has been designed for the present-day world. One consequence of this is that lexical items can be attributed to incorrect semantic categories; that is to say, "incorrect" according to the worldview contemporaneous with the text. For example, the word *cousin* is ascribed to **S4** Kin, but in Shakespeare's period is simply denoted a 'friend', and should therefore be in **S3.1** Relationship*:* General. Some of these specific problems have been corrected by the historical lexicon. However, one might go on to argue that the very structure of the semantic categories would be somewhat different. For example, the present-day view of intimacy implying a sexual relationship is reflected in the category: **S3.2** Relationship: Intimate/sexual. However, the sexual mores of the Elizabethan world were different, with the consequence that intimacy was possible without such strong sexual implications. The word *lover*, ascribed to **S3.2** by SEMTAG, is a case in point, as in this period it had the sense of 'friends' or 'intimates', but no strong implications of sexual relations. A counter argument in support of USAS might be that reading Shakespeare through the prism of the present-day worldview is the majority experience today. Few of us — if any — are sufficiently steeped in Elizabethan social history in order to be able to transcend our own milieu. Nevertheless, cautious checking and interpretation of the results is required.

Table 5 displays the semantic categories that are key in Romeo's speech. Categories with fewer instances than 15 (legitimate) lexical tokens — a figure derived after the scrutiny of every member of every category — are noticeably less robust and well motivated than the others. For example, in Education in general (P1), the connection between *philosophy* and *school* is rather tenuous (even more so when these words are read in context). The notion of a semantic category is obviously more abstract than that of a lexical item. Such categories need a certain weight of both lexical types and tokens before commonalities can be clearly seen. Also,

Table 5. Romeo's semantic categories rank-ordered for positive keyness (i.e. relatively unusual over-use) (keywords, as listed in Table 1, are emboldened)

Semantic category, including the tag code and frequency	Items within the category (and their raw frequencies) up to a maximum of ten types if they are available (excluding clear tagging errors in square brackets)
Relationship: Intimate/sexual (S3.2) (48)	***love*** (34), *kiss* (5), *lovers* (3), *kisses* (2), *paramour* (1), *wantons* (1), *chastity* (1), *in love* (1)
Liking (E2+) (38)	***love*** (15), ***dear*** (13),[12] *loving* (3), *precious* (2), *like* (1), *doting* (1), *amorous* (1), [*revels* (1)], *loves* (1)
Colour and colour patterns (O4.3) (33)	*light* (6), *bright* (4), *pale* (3), *dark* (3), *green* (2), *stained* (2), *black* (2), *golden* (1), *white* (1), *crimson* (1)
Education in general (P1) (9)	*teach* (3), [*course* (2)], *philosophy* (2), *school* (1), *schoolboys* (1)
Business: Selling (I2.2) (19)	*sell* (4), [*bid* (4)], *shop* (2), *hire* (2), *buy* (1), *sold* (1), [*stands* (1)], [*bade* (1)], [*stand* (1)], [*store* (1)], *merchandise* (1)
Thought, belief (X2.1) (26)	*think* (7), *feel* (3), *devise* (2), *believe* (2), [*take thence* (1)], *thinking* (1), *thought* (1), *engrossing* (1), *dreamt* (1), [*found* (1)], *in thine eyes* (1), *in mind* (1)
Affect: Cause/Connected (A2.2) (20)	[*hence* (7)], *reason* (2), [*spurs* (2)], *depend* (1), *for fear of* (1), *provoke* (1), *excuse* (1), *effect* (1), *consequence* (1), *to do with* (1), *appertaining* (1), *prompt* (1)
Avarice (S1.2.2+) (7)	*envious* (3), [*mean* (1)], *tempt* (1), *jealous* (1), *sparing* (1)
The universe (W1) (21)	*world* (8), [*word* (6)], ***stars*** (5), *moon* (2)
Money: Affluence (I1.1+) (7)	***rich*** (7)

semantic categories with fewer than 15 tokens in total tend to be localised. For example, regarding the category Business: Selling (I2.2), all instances of *sell*, *buy*, *sold* and *shop* occur in Act V. scene i, where Romeo purchases poison. Consequently, my discussion will focus on categories which have more than 15 tokens.

As can be seen from Table 5, very few lexical items constituting the semantic categories are also keywords. This again suggests that Rayson (2004) was right to say that semantic categories group together lexical items whose low frequencies will prevent them from being identified as key by themselves. The first two categories, Relationship: Intimate/sexual (S3.2) and Liking (E2+), are, of course, very closely related categories, and some argue that 'liking' stands in a metonymic relationship with 'love' (see Barcelona Sánchez 1995: 675). Taken as a whole, the identification of these categories as most key is very well motivated, a predictable finding that confirms Romeo's role as the lover of the play.[13] The third top-most category, Colour and colour patterns (O4.3), is much less predictable. Sometimes Romeo is describing literal light: *But, soft! what light through yonder window breaks?* (II.ii). But, more often, the terms are used metaphorically, as in for example: *More light and light; more dark and dark our woes (III.v); Be not her maid, since she is envious; Her vestal livery is but sick and green* (II.ii); *Death […] Hath had no power yet upon thy beauty: Thou art not conquer'd; beauty's ensign yet Is crimson in thy lips and in thy cheeks, And death's pale flag is not advanced there* (V.iii). These are all fairly conventional metaphors: light/dark for happiness/unhappiness, greenness for envy, and redness/whiteness for life/death. The semantic tagger cannot yet distinguish between literal and metaphorical meanings.[14] Nevertheless, some of the semantic categories picked out reveal metaphorical patterns (see Archer et al. 2009, for a more extensive demonstration of this). Metaphor is often used to express emotions in more concrete terms, and Romeo's colour terms often do just that. Note that this contributes to his characterisation: the Nurse gives it to you straight with the repetition of expressions like *O woeful day*, whereas Romeo uses metaphor: *more dark and dark our woes*. This helps account for Romeo as the more complex character playing a more central role in the play, and also, possibly, as a character of higher status. The fact that the category Thought, belief (X2.1) is also key is consistent with the idea that he is more complex: he is a reflective character who reveals his thoughts to the audience.

Table 6 displays the semantic categories that are key in the Nurse's speech. A rather larger number of keywords appear in the Nurse's semantic categories compared with either Romeo or Mercutio. In part, this may be due to the fact that the Nurse's speech is characterised by a high degree of repetition.[15] Importantly, however, not as many semantic categories include keywords as it might seem at first sight. Two discourse markers, *faith* and *warrant*, incorrectly appear in the categories Worry, concern, confident (E6+) and Obligation and necessity (S6+), respectively; their historical meanings, however, place them in Discourse Bin (Z4) (the category comprised of discourse markers). Similarly, *well* in Evaluation: Good/bad (A5.1+) is also nearly always a discourse marker. Had these items been correctly tagged it is somewhat doubtful whether the categories in which they currently appear would still be key. Furthermore, the categories People: Female (S2.1) and Power, organizing (S7.1+) overlap. The vocatives *lady* and *madam* clearly also have implications of power, and thus should also appear under that category. Conversely, *mistress* also has the sense of female, and so could also appear under that category.

A characteristic of both the items in People: Female (S2.1) and Power, organizing (S7.1+) is that they contain vocatives. They help construct the kind of social

Table 6. Nurse’s semantic categories rank-ordered for positive keyness (i.e. relatively unusual over-use) (keywords, as listed in Table 1, are emboldened)

Semantic category, including the tag code and frequency	**Items within the category (and their raw frequencies) up to a maximum of ten types if they are available (excluding clear tagging errors in square brackets)**
People: Female (S2.1) (31)	***lady*** (15), ***madam*** (10), *girl* (2), *gentlewoman* (2), *women* (1), *woman* (1)
Discourse Bin (Z4) (48)	***God*** (11), ***ah*** (6), ***ay*** (5), *nay* (4), *no* (2), *alack* (2), *as I said* (2), [*ne’er* (2)], [*forget it* (2)], *you know* (2), *fie* (2), *as they say* (2)
Time: Period (T1.3) (41)	***day*** (18), *night* (5), *years* (3), *days* (3), [*stinted* (2)], *awhile* (2), *year* (2), [*second* (1)], *for a week* (1), *nights* (1), *hour* (1), [*mar* (1)], *afternoon* (1)
Time: Old, new and young; age (T3) (6)	*age* (2), *fourteen* (2), *twelve year old* (1), *eleven* (1)
Happy/sad: Happy (E4.1-) (27)	***woeful*** (6), *alas* (3), *lamentable* (3), *weeps* (2), *crying* (2), *piteous* (2), *pitiful* (1), *woe* (1), *sorrows* (1), *weeping* (1)
Entirety; maximum (N5.1+) (25)	*all* (16), *any* (5), *every* (2), *full* (1), [*gross* (1)]
Worry, concern, confident (E6+) (9)	[***faith*** (8)], *confidence* (1)
Obligation and necessity (S6+) (23)	[***warrant*** (7)], [*should* (6)], *must* (6), *needs* (3), *need* (1)
Power, organizing (S7.1+) (31)	***sir*** (13), ***lord*** (10), *mistress* (4), [*beats* (1)], *nobleman* (1), [*say* (1)], *lead* (1)
Being (A3+) (90)	*is* (28), *‘s* (20), *be* (15), *were* (8), *was* (7), *are* (6), *am* (6)
Generally kinds, groups, examples (A4.1) (9)	*case* (4), *kind* (2) [*side* (1)], [*come to* (1)], [*coming to* (1)]
Evaluation: Good/bad (A5.1+) (21)	[***well*** (10)], *good* (8), *excels* (2), *great* (1)

network of which the Nurse is a part: she interacts with women such as Juliet and Lady Capulet, and also powerful individuals in the family such as Capulet. Items of this kind constitute a relatively restricted set that can potentially be used frequently, and so it is not surprising that there is overlap with the keyword results. The same can be said of the second most key semantic category, Discourse Bin (Z4), which is well-stocked with keywords — discourse markers were revealed in my keyword analysis as a strong feature of the Nurse’s style. In contrast, the category Being (A3+) contains no keywords at all (although we can note that the keyword *he’s* in Table 1 also contains part of the verb *to be*). This is despite the fact that it is populated by a restricted set of word types. The category is well motivated:

it reflects the Nurse's role as the irrepressible commentator in the play. She freely gives her opinion (*my mistress is the sweetest lady-Lord*, *Paris is the properer man*, *He is not the flower of courtesy*), states what is happening (*Your lady mother is coming to your chamber*), predicts what will happen (*Come Lammas-eve at night shall she be fourteen*, *This afternoon, sir? well, she shall be there*, *your Romeo will be here to-night*) and expresses doubts (*If you be he, sir, I desire some confidence*, *Marry, that, I think, be young Petruchio*). The category Entirety; maximum (N5.1+) also contains no keywords, but seems well motivated too. It reflects the Nurse's tendency to dramatize events she narrates, as in the following example: *A piteous corse, a bloody piteous corse; Pale, pale as ashes, all bedaub'd in blood, All in gore blood; I swounded at the sight*. (III.ii)

Table 7 displays the semantic categories that are key in Mercutio's speech.

Table 7. Mercutio's semantic categories rank-ordered for positive keyness (i.e. relatively unusual over-use) (keywords, as listed in Table 1, are emboldened)

Semantic category, including the tag code and frequency	**Items within the category (and their raw frequencies) up to a maximum of ten types if they are available (excluding clear tagging errors in square brackets)**
Living creatures generally (L2) (34)	***hare*** (5), [*bawd* (3)], *egg* (2), *dog* (2), *cats* (2), *mouse* (2), *wings* (2), *flies* (1), *herring* (1), *goose* (1), *rat* (1)
Grammatical bin (Z5) (606)	***the*** (82), ***a*** (82), ***of*** (58), *and* (52), *to* (31), *in* (28), *for* (26), *with* (21), *'s* (*s*-genitive) (16), *as* (16)
Shape (O4.4) (9)	*straight* (4), [*row* (1)], *sharp* (1), *round* (1), *shape* (1), *circle* (1)
Food (F1) (14)	*meat* (3), [*hams* (1)], [*sauce* (1)], [*peppered* (1)], *pie* (1), *dinner* (1), *nuts* (1), [*grub* (1)], *bakes* (1), *fruit* (1), *pear* (1), *butcher* (1)
Clothes and personal belongings (B5) (12)	*wearing* (2), *worn* (1), *wear* (1), *livery* (1), *tailor* (1), *doublet* (1), *shoes* (1), *ribbon* (1), [*suit* (10)], *collars* (1), *button* (1)
Anatomy and physiology (B1) (64)	***eye*** (4), *face* (3), *asleep* (3), *hair* (3), *ears* (3), *eyes* (3), *ear* (3), *bosom* (2), *head* (2), *nose* (2), *flesh* (2)
Arts and crafts (C1) (15)	[*art* (10)], [*draws* (1)], [*joiner* (1)], *coach-makers* (1), [*draw* (1)], [*drawn* (1)]
Health and disease (B2-) (14)	*plague* (3), *scratch* (3), *faints* (1), *faint* (1), *hurt* (1), *blisters* (1), *plagues* (1), *mad* (1), [*black eye* (1)], *pox* (1)

Mercutio speaks fewer words than Romeo or the Nurse, and so it is not surprising that only three semantic categories are above my cut-off point of 15. All three categories, Living creatures generally (L2), Grammatical bin (Z5) and Anatomy and physiology (B1), contain at least one keyword. The Grammatical bin (Z5), like the Discourse Bin (Z4) for the Nurse, contains three keywords as the most frequent items. For Living creatures generally (L2) and Anatomy and physiology (B1), the semantic analysis suggests that the keywords *hare* and *eye* are part of larger semantic categories, and that those categories are a significant, distinctive feature of Mercutio's style. Together, they suggest a focus on the physical and animate, as can be seen in the following example (words relevant to the two categories under discussion are underlined):

> thou wilt quarrel with a man that hath a hair more or a hair less in his beard than thou hast. Thou wilt quarrel with a man for cracking nuts, having no other reason but because thou hast hazel eyes. What eye, but such an eye, would spy out such a quarrel? Thy head is as full of quarrels as an egg is full of meat, and yet thy head hath been beaten as addle as an egg for quarrelling. Thou hast quarrelled with a man for coughing in the street, because he hath wakened thy dog that hath lain asleep in the sun. (III.i)

The semantic analyses have very marginally less overlap with keywords compared with part-of-speech analyses. It is only for the analysis of Mercutio, characterised by textual keywords, that we find a keyword in all key semantic categories (above the threshold set for a robust category). Unlike the part-of-speech categories, semantic categories generally contained a range of types, something which is to be expected, given that semantic categories operate at a higher level of abstraction. Like the part-of-speech categories, semantic categories were usually dominated by one or two very frequent items, and this may partly explain why they overlap with keywords. Where the key analysis of semantic tags has an advantage over grammatical annotation is that the semantic categories revealed as key and not consisting of keywords are also well motivated and difficult to predict. Romeo's metaphorical colour patterns, for example, or the items relating to 'being' for the Nurse are illuminating and fit our intuitions about their characters. Given that the analyses here have engaged relatively small datasets, it may be the case that these more abstract semantic categories can be more effectively revealed in larger sets of data, and, indeed, early work suggests that this is the case. Archer et al. (2009) explore Shakespeare's 'love tragedies' and 'love comedies', and show how key semantic categories tap into metaphorical patterns. Overall, identifying what is semantically key has potential, and potential beyond what a keyword analysis can reveal, though very careful checking of the results is required, as well as the institution of thresholds for robustness, due to the relatively low reliability of the tagging.

8. Conclusions

A conclusion that a keyword analysis simply provides evidence for what one might have predicted — for example, establishing that Romeo is all about love — would not be accurate. A keyword analysis has two other valuable aspects compared with traditional qualitative analyses. Firstly, it can reveal features that are less obvious and therefore less easily observable (and thus often overlooked) but which cannot safely be assumed to have a negligible effect. For example, Juliet's keywords *if*, *yet*, *or*, *would* and *be* (mostly subjunctive) create a grammatical style that can be interpreted as evidence of the anxieties we are likely to understand her as experiencing in the play. It is no surprise that researchers in critical discourse analysis have deployed keyword analyses in order to reveal hidden, ideologically-driven discourses (see, for example, Gabrielatos & Baker's 2008 analysis of refugee discourse). Secondly, it can reveal lexical and grammatical patterns without reliance on intuitions about either which parts of the text to focus on or what the relevant dimensions or features are. Many other quantitative methods — multi-dimensional analysis, for example — involve a priori decisions about what to count, but keyword analysis does not.

I have argued that the notion of a keyword has a history of at least 50 years, and is closely connected with a notion of style, and, more specifically, the notion of style marker. Keyword analyses are not new either, though the ease and speed at which one can perform them, thanks to programs such as *WordSmith Tools* and *WMatrix,* is. However, the recent explosion in studies involving keyword analyses has not been matched by sensitivity to the procedures they involve. This paper reviewed some of the procedural decisions (e.g. the choice of reference corpus, the frequency cut-off, type of statistical test, probability value). A general issue for each keyword analysis is that each individual study tends to use its own settings and sometimes different reference corpora. This, of course, raises issues of comparability. However, in practice, studies tend to focus their discussion on the most key keywords, which would most likely arise in the context of various normally used settings. Regarding keyword analysis results, I proposed that they can be categorised according to three (Halliday-derived) functional emphases: ideational (as illustrated by Romeo), textual (as illustrated by Mercutio) or interpersonal (as illustrated by the Nurse).

What is to be gained from extending a keyness analysis to part-of-speech categories or semantic categories? Tables 8 and 9 display the number of key part-of-speech and semantic categories that arose and how many of those were dominated by words which were also retrieved in the keyword analysis (dominated here means the category item with the most tokens). (For semantic categories, only categories with 15 or above legitimate tokens are counted).

Table 8. The number of key part-of-speech categories dominated by one or two keywords

Character (and Table)	Number of key part-of-speech categories	Number of key categories dominated by one or two keywords
Romeo (Table 2)	6	4
Nurse (Table 3)	5	5
Mercutio (Table 4)	5	3
Total	16	12

Table 9. The number of key semantic categories dominated by one or two keywords

Character (and Table)	Number of key semantic categories	Number of key categories dominated by a keyword
Romeo (Table 5)	5	2
Nurse (Table 6)	7	5
Mercutio (Table 7)	3	3
Total	15	10

In percentage terms, 75% of the part-of-speech categories are dominated by one or two words that also occur as keywords, and 66.6% of the semantic categories are dominated by one or two words that also occur as keywords. So, whilst the study discussed here clearly needs to be replicated on larger sets of data, it seems to be the case that generally we can 'trust the text' (Sinclair 2004): a straight keyword analysis revealed most of the conclusions. On the face of it, the difference between 75% and 66.6% seems inconsequential. It is worth noting, however, that the analyses for Romeo have rather less overlap with the keyword analysis, the relevant figures being 66% for part-of-speech categories and 40% for semantic categories. When keywords are dominated by ideational keywords, capturing the 'aboutness' of the text, the part-of-speech and particularly the semantic keyness analyses have much more of a contribution to make, moving the analysis beyond what is revealed in the keywords. The probable reason for this is that more grammatical items, and also discourse markers, are dominated by a relatively restricted range of types each with frequent tokens. Thus, if categories including such items are identified as key in the part-of-speech or semantic analyses, as in the cases of the Nurse and Mercutio, then it is highly likely that they will also appear in the keyword analysis.

As pointed out at the beginning of this paper, Rayson (2004, 2008) argues for conducting key part-of-speech and semantic analyses in addition to keyword analyses, and the analyses of this paper suggest that he has some justification. Firstly, he argues that there are fewer categories for the researcher to grapple with. Whilst it may well be the case that fewer keywords can be produced by tweaking the program settings, it is also the case that by conducting key part-of-speech and semantic analyses one can get clues as to patterns that exist in a large set of keywords. Secondly, he argues part-of-speech and semantic categories can group lower frequency words which might not appear as keywords individually and could thus be overlooked. This has been confirmed in the analyses of this paper, examples of which include general adjectives and (metaphorical) colour terms for Romeo, plural common nouns for Mercutio and items relating to 'being' for the Nurse. None of these are easily predictable, but all seem well-motivated upon closer inspection. Note, however, that the additional contribution provided by these analyses is more specific than Rayson suggests: it only pertains to more lexical, more ideational categories. This is consistent with my argument at the end of the previous paragraph. It is not surprising, then, that we are seeing current research efforts applying semantic categories and keyness analysis in order to reveal metaphorical patterns in various discourses (see, for example, Archer et al. 2009; Koller et al. 2008).

To conclude, it is worth flagging up the limitation — danger even — emphasized by Baker (2004) that a keyword analysis, and in fact any keyness analysis, encourages the research to focus on differences at the expense of similarities. In terms of *Romeo and Juliet*, revealing one character's statistically-based similarities with another could indeed be enlightening.

Notes

* I thank the two anonymous reviewers for their comments and particularly Michaela Mahlberg for her feedback, advice and patience. Of course, I am entirely responsible for the final paper.

[1] This section of the paper is based on Culpeper (2002).

[2] Of course, this is not the only program that can do this kind of analysis. For example, *AntConc* and *WMatrix* can also do it. There is some variation, however, with respect to what parameters they have available for tweaking.

[3] *WordSmith Tools* lists negative keywords in ascending order of keyness. Further, readers may note slight differences between this table and Table 3 in Culpeper (2002:19). I repeated the keyword analysis for this table at the same time as I did the other analyses in this paper. All settings were the same. I cannot explain why there are very slight differences, but these differences are not meaningful and, in particular, do not affect any of the claims I make about the results.

[4] Scott's example of *already* as a stylistic keyword is a borderline open/closed class word. Adverbs are a varied and thus problematic category, which is considered open class by some linguists and closed by others.

[5] In fact, items such as *if* and *or* relate very clearly to the "logical" subcomponent of Halliday's ideational metafunction — they are not devoid of "content".

[6] More information about CLAWS tagger can be found at: http://www.comp.lancs.ac.uk/ucrel/claws. A free web-based tagging service is available (though with a few restrictions).

[7] The tagger separates possessive pre-nominal pronouns (e.g. *your*, *our*) (APPGE), functioning as a determiner, from nominal possessive personal pronouns (e.g. *yours*, *ours*) (PPGE), the category under consideration here. The single appearance of *his* as PPGE is an oddity; all the other instances are correctly tagged. Potentially more problematic is the fact that in this period *mine* and *thine* could perform both functions. *My/mine* and *thy/thine* could both be pre-nominal determiners, the –n forms being preferred if the following noun began with a vowel. However, the tagger correctly identifies instances like *I have been feasting with mine enemy* as the pre-nominal pronoun, and instances like *heaped like mine, and that…* as the nominal pronoun. The only error occurs with the phrase *mine own* (4 occurrences), which is treated as a nominal pronoun. Generally the tagger is accurate. Furthermore, I note that the category APPGE also appears in Romeo's key parts of speech list (though ranked 19th and thus well below the significance level required). These tagging inaccuracies would make very little difference to the analytical conclusions.

[8] All of these, except one instance, occur within vocative constructions.

[9] Approximately half of these occur within vocative constructions.

[10] In the keyword analysis, the word-form *I* achieved a log-likelihood value of 5.52, and the word-form *than* achieved a critical value of 6.18, both of which are below the critical value of 6.63 for $p < 0.01$. The possible reason why these single items are key constituting grammatical categories but not key as words is because the more items that constitute a category the less easy it is for differences in the frequency of that category to emerge, as differences relating to a specific member of that category could be averaged out by other members. So, in the context of differences between words, an item might not be key, but in the context of differences between grammatical categories that are usually comprised of more than one word-form they may be key.

[11] The instances of *o* are of the interjection, *a* the definite article, and *I* the pronoun. Only *r* in *R is for the-No* is accurately tagged as a single letter of the alphabet. The tagger was perhaps confused by items (e.g. *piteous*, *woeful*, *courteous*) to the right. The important point to note is that the instances of *o* would further bolster the interjection category later in the table.

[12] Most of these are not in fact an element in a vocative.

[13] The appearance of *love* in both categories is not an error. As I pointed out, the semantic tagger also uses POS information. Verbal usages of *love* are generally assigned to Liking (E2+).

[14] The use of the semantic tagger for metaphor identification is being investigated by Elena Semino and Veronika Koller, amongst others, at Lancaster University.

[15] Culpeper (2001:188–190) undertook a type-token ratio analysis and compared the Nurse with Capulet and Mercutio. The Nurse used significantly more repetition.

References

Afida, M. A. 2007. "Semantic fields of problem in business English: Malaysian and British journalistic business texts". *Corpora*, 2 (2), 211–239.

Archer, D. (Ed.) 2009. *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*. London: Ashgate.

Archer, D., Culpeper, J. & Rayson, P. 2009. "Love — 'a familiar of a devil'? An exploration of key domains in Shakespeare's comedies and tragedies". In D. Archer (Ed.), *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*. London: Ashgate.

Archer, D., McEnery, T., Rayson, P. & Hardie, A. 2003. "Developing an automated semantic analysis system for Early Modern English". In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference. UCREL technical papers number 16*. Lancaster University, Lancaster: UCREL, 22–31.

Archer, D. & Rayson, P. 2004. "Using an historical semantic tagger as a diagnostic tool for variation in spelling". Presented at the *Thirteenth International Conference on English Historical Linguistics*, University of Vienna, Austria.

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

Barcelona Sánchez, A. 1995. "Metaphorical models of romantic love in *Romeo and Juliet*", *Journal of Pragmatics*, 24, 667–688.

Berber Sardinha, T. 1999: online. "Using KeyWords in text analysis: Practical aspects". *DIRECT Working Papers* 42*, São Paulo and Liverpool*. Available at: http://www2.lael.pucsp.br/direct/DirectPapers42.pdf (accessed September 2008).

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G. N., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Brinton, L. & Traugott, E. C. 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

Burrows, J. F. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Craig, W. J. 1914. *William Shakespeare (1564–1616). The Oxford Shakespeare*. Oxford: Oxford University Press.

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 U. Melander-Marttala, C. Ostman & M. Kytö (Eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium*, *Association Suédoise de Linguistique Appliquée (ASLA),* 15. Uppsala: Universitetstryckeriet, 11–30. (Also available at: http://www.lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm)

Enkvist, N. E. 1964. "On defining style". In N. E. Enkvist, J. Spencer & M. Gregory (Eds.), *Linguistics and Style*. Oxford: Oxford University Press, 1–56.

Gabrielatos, C. & Baker, P. 2008. "Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005". Journal of English Linguistics, 36 (1), 5–38.

Garside, R. 1987. "The CLAWS word-tagging system". In R. Garside, G. Leech & G. Sampson (Eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman, 30–56.

Guiraud, P. 1954 [1970]. *Les Caractères Statistiques du Vocabulaire*. Pages 64–7 reprinted in: P. Guiraud & P. Kuentz (Eds.), *La Stylistique Lectures*. Paris: Klincksieck, 222– 224.

Halliday, M. A. K. 1973. *Explorations in the Functions of Language*. London: Edward Arnold.

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: Edward Arnold.

Hoover, D. L. 1999. *Language and Style in The Inheritors*. Lanham, Maryland: University Press of America.

Johnson, S., Culpeper, J. & Suhr, S. 2003. "From 'politically correct councillors' to 'Blairite nonsense': Discourses of political correctness in three British newspapers". *Discourse and Society*, 14 (1), 28–47.

Jones, M., Rayson, P. & Leech, G. 2004. "Key category analysis of a spoken corpus for EAP". Presented at *The 2nd Inter-Varietal Applied Corpus Studies* (IVACS) *International Conference on 'Analyzing Discourse in Context'*, The Graduate School of Education, Queen's University, Belfast, Northern Ireland, 25–26 June.

Koller, V., Hardie, A., Rayson, P. & Semino, E. 2008: online. "Using a semantic annotation tool for the analysis of metaphor in discourse". *Metaphorik.de* Available at: http://www. metaphorik.de/15/.

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. (Also available at: http://www.comp.lancs.ac.uk/computing/research/ucrel/papers/coling.html).

Leech, G. N. & Short, M. 2007 [1981]. *Style in Fiction*. London: Longman.

Longman Lexicon of Contemporary English. 1981. McArthur, T. London: Longman.

Phillips, M. 1989. *Lexical Structure of Text*. (Discourse Analysis Monographs, 12). Birmingham: University of Birmingham.

Piao, S. S. L., Rayson, P., Archer, D. & McEnery, T. 2004. "Evaluating lexical resources for a semantic tagger". In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, May 2004, Lisbon, Portugal, Volume II, 499–502.

Rayson, P. 2003. *Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison*. Ph.D. thesis, Lancaster University.

Rayson, P. 2004. "Keywords are not enough". Invited talk for JAECS (Japan Association for English Corpus Studies) at Chuo University, Tokyo, Japan. http://www.comp.lancs.ac.uk/computing/ users/paul/public.html

Rayson, P. 2005. *WMatrix: A Web-based Corpus Processing Environment*. Computing Department, Lancaster: Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix/.

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

Rayson, P. & Wilson, A. 1996. "The ACAMRIT semantic tagging system: progress report". In L. J. Evett & T. G. Rose (Eds.) *Language Engineering for Document Analysis and Recognition*, LEDAR, AISB96 Workshop Proceedings, Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK, 13–20.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. 2004. "The UCREL semantic analysis sytem". 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)*, 25th May 2004, Lisbon, Portugal. Paris: European Language Resources Association, 7–12.

Rayson, P., Archer, D. & Smith, N. 2005: online. "VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora". In *Proceedings from the Corpus Linguistics Conference Series On-line E-journal*, 1 (1). Available at: http://www.corpus.bham.ac.uk/PCLC/ and http://eprints.comp.lancs.ac.uk/1157/ (Accessed September 2008).

Scott, M. R. 1997. "PC analysis of key words — and key key words". *System*, 25 (2), 233–45.

Scott, M. R. 2000. "Focusing on the text and its key words". In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective*, Volume 2. Frankfurt: Peter Lang, 103–122.

Scott, M.R. 2008. *WordSmith Tools Help Manual*. Version 5.0. Liverpool: Lexical Analysis Software.

Scott, M. R. & Tribble C. 2006. *Key Words and Corpus Analysis in Language Education*. Amsterdam & Philadelphia: John Benjamins.

Sinclair, J. McH. 2004. (Edited with R. Carter) *Trust the Text: Language, Corpus and Discourse*. London & New York: Routledge.

Tribble, C. 2000. "Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals". In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang, 75–90.

Williams, R. 1976. *Keywords: A Vocabulary of Culture and Society*. London: Fontana.

Wilson, A. & P. Rayson 1993. "The automatic content analysis of spoken discourse". In C. Souter & E. S. Atwell (Eds.), *Corpus-based Computational Linguistics*. Amsterdam: Rodopi, 215–216.

Wynne, M. 2006. "Stylistics: Corpus approaches". In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd ed.). Oxford: Elsevier, 223–25.

Xiao, R. & McEnery, T. 2005. "Two approaches to genre analysis: Three genres in Modern American English". *Journal of English Linguistics*, 33 (1), 62–82.

Author's address

Jonathan Culpeper
Department of Linguistics and English Language
Bowland College
Lancaster University
Bailrigg
Lancaster
LA1 4YT
U.K.

j.culpeper@lancaster.ac.uk