[This paper appeared in print as: Culpeper, J. (2002) 'Computers, language and characterisation: An Analysis of six characters in Romeo and Juliet'. In: U. Melander-Marttala, C. Ostman and Merja Kyto (eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium*, Association Suedoise de Linguistique Appliquee (ASLA), 15. Universitetstryckeriet: Uppsala, pp.11-30. Note that the pagination here follows that of this publication.]

Computers, language and characterisation: An analysis of six characters in *Romeo and Juliet*

Jonathan Culpeper, Lancaster University

1. The problem

For the last dozen years or so I have been investigating how language is used to generate particular impressions of characters – or, indeed, people – in the heads of readers or audiences.[1] I found two areas, lexis and grammar, particularly problematic. There is a large literature on how lexical and grammatical items constitute particular registers and dialects, which may correlate with particular social groups. What we know less about is how such items reflect the speech styles of particular personalities. Intuitively, it is reasonable to suggest that lexis plays a significant role. For example, the tendency to use formal lexis may – context permitting – give the impression that someone is rather aloof or pompous; informal lexis that someone is 'down to earth'. But what the relevant lexical dimensions are and how one goes about revealing them in an analysis has been largely overlooked. The situation for grammatical features is similar. For example, Scherer, in his review of personality markers in speech, is surprised that 'there is no systematic research on personality differences in cognitive processing and the complexity of syntactic structure' (1979: 170). One particular problem is that patterns created by grammatical features are often unconsciously observed. Page, for example, examining speech in the English novel, states: 'Grammar and syntax are, apart from the most obvious differences, less readily absorbed by the casual listener, and are used relatively little by writers' (1988: 57). Similarly, Blake (1983), considering Shakespeare, concludes that syntactic differences between characters' speech are less important than other aspects because they are less likely to be noticed: they are 'more subtle than marked features of vocabulary or dialect and can readily be overlooked, particularly in the theatre' (1983: 28). The important point, I would argue, is that, although these features are less obvious and therefore less easily observable, this does not mean that we can safely assume that they have a negligible effect on our impressions. It might be the case that the accumulative effect of lexical or grammatical features is decisive in shaping an impression of character.

The broad aim of this paper is to show how the study of an important area within 'stylistics', namely characterisation, can benefit from an empirical approach, specifically, a methodology for identifying what might be the 'key' words of a text. Such an approach can reveal significant lexical and grammatical patterns without reliance on speculations about what the relevant dimensions are. I shall start by arguing that the notion of 'key' words relates to what Enkvist (1964, 1973) called 'style-markers'. Using the *Keywords* facility in Mike Scott's *WordSmith Tools* (1999), a computer program that does the kind of analysis required of Enkvist's definition, I demonstrate this with an analysis of character speech in Shakespeare's *Romeo and Juliet.* Having generated a list of keywords for the main characters, I examine the function and context of the keywords, in order to validate and account for the results. I conclude by noting both further possibilities and limitations for stylistics and corpus linguistics in general, and keywords analysis in particular.

2. Style, style-markers and style-reminders

Nils Erik Enkvist's (1964, 1973) definition of style lends itself well to statistical analysis:

> 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's, 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)

Style, then, is a matter of 'frequencies', 'probabilities' and 'norms'. He goes on to offer the following definition of 'style markers':

> 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, then, are words whose frequencies differ significantly from their frequencies in a norm. As we shall see in the next section, this is precisely the principle used to identify 'keywords'. One other concept worth drawing attention to here is that of 'style reminders'. It is reasonable to suppose that style markers, by virtue of the fact that they are 'contextually bound', acquire stylistic meaning (cf. Leech 1981: 14-15) over time. Thus, a single occurrence of a style reminder may convey certain stylistic associations, regardless of the context in which it appears. For example, my students have always readily suggested that the word 'caste' evokes the biblical register, whilst 'chuck' evokes a colloquial register (with 'throw' the rather more neutral term), even when I give them these words without context. Clearly, the unusual frequency with which 'caste' occurs in the biblical register and 'chuck' in the colloquial register (a matter of style markers), has set up the conditions by which they can develop into style reminders. This paper will focus on statistically defined style markers.

3. Keywords via *WordSmith Tools* (1999)

Keywords, here, are 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). The term keywords can be seen as another term for style markers. In fact, Enkvist (1973: 132-3) alludes to Pierre Guiraud's '*mots-clès*'. I shall adopt the term keywords in this paper, because in doing so I can make a clear link to developments in corpus linguistics. Specifically, the notion of keywords has been developed and popularised by Mike Scott, through the *KeyWords* facility of his program *WordSmith Tools* (1999), a program designed for the analysis of corpora. This program performs the kind of statistical analysis required to identify keywords. It conducts a statistical comparison between the words of a corpus (or wordlist) and a bigger reference corpus, in order to identify words that are unusually frequent or unusually infrequent. According to Scott (1999: Help Menu):

> 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', then, is a matter of being statistically unusual. The statistical operations involved here – a cross tabulation, a chi-square significance test – are amongst the most basic in statistics, and common in the world of corpus linguistics. However, to apply such operations manually would be extremely time-consuming. The chief benefit of such a program is that one can load in the relevant texts and get results within minutes.

4. Preparing the text

The fact that I had chosen to analyse a Shakespearean play-text raised a couple of issues that had to be resolved before analysis could start. The first issue related to the choice of edition. One possibility was to download an electronic version of *Romeo and Juliet*, the First Folio, from the Oxford Text Archive. Although this has the merit of being a more 'original' text, it also contains much spelling variation. Spelling variation is perhaps the greatest obstacle in the statistical manipulation of historical texts. Computers focus on word forms: 'sweete', for example, would not be counted along with 'sweet'. For this reason, I downloaded *The Oxford Shakespeare* (1914, edited by W. J. Craig) edition of the play, complete with modern standardised spelling, from the web. The second issue relates to the fact that Shakespearean plays consist of dialogue. Clearly, there has to be a way of enabling the computer to distinguish between the speech of different characters. To that end, I added a simple tagging system, consisting of a switch-on tag based on the first three letters of the character's name, and a switch-off tag based on a backslash and the first three letters of the character's name.[2] A sample of the tagged text follows:

Enter SAMPSON and GREGORY, armed with swords and bucklers.
<SAM>Gregory, o' my word, we'll not carry coals.<\SAM>
<GRE>No, for then we should be colliers.<\GRE>
<SAM>I mean, an we be in choler, we'll draw.<\SAM>
<GRE>Ay, while you live, draw your neck out o' the collar.<\GRE>
<SAM>I strike quickly, being moved.<\SAM>
<GRE>But thou art not quickly moved to strike.<\GRE>
<SAM>A dog of the house of Montague moves me.<\SAM>

As can be seen from this sample, such a tagging system also enables one to exclude non-speech material, such as stage directions. *WordSmith Tools* (1999) is well suited to operating with such a tagging system, since it allows one to examine text between a specific set of tags or to exclude text between a specific set of tags from one's examination.

5. Selecting characters, comparators and parameters

The major criterion determining which and how many characters I investigated was how many words they spoke. Table 1 displays in rank order the total number of words spoken by seven characters in *Romeo and Juliet*.

Table 1. The total number of words spoken by seven characters in *Romeo and Juliet*

Character	*Total no. of words spoken*
Romeo	5,031
Juliet	4,564
Friar Lawrence	2,901
Nurse	2,369
Capulet	2,292
Mercutio	2,254
Benvolio	1,293

It is clear from this table that a cut-off point presents itself after Mercutio, since the word count for the next character, Benvolio, drops by nearly a 1000 words.

In any keywords analysis, the choice of data for comparison (the reference list) is crucial. There is no magic formula for making this decision. Some clues are provided in the quotations from Enkvist in section 2 above. Clearly, a set of data which has no relationship with the data to be examined is unlikely to reveal interesting results (cf. Enkvist's comparison of a Shakespearean sonnet with a telephone directory). An important factor will be one's research goal. In my

case, comparing each of the six characters with all of the other characters in the play, seemed to be the obvious choice.[3] Characters are partly shaped by their context. Thus, it makes little sense to compare, say, the characters of *Romeo and Juliet* with the characters of *Macbeth* or *Anthony and Cleopatra*, since the fictional worlds of Italy, Scotland and Egypt provide very different contextual influences. Furthermore, characters, like people, are partly perceived in terms of whom they interact with. Indeed, linguists have argued that interaction itself can reveal personality. Brown and Levinson put it thus: 'an understanding of the significant dimensions on which interaction varies should provide insights into the dimensions on which personality is built, as well as social relationships' (1987: 232).

The *Keywords* program (within *WordSmith Tools*) allows the user to set various parameters. I will briefly note my settings here. I set the minimum frequency for a word to be considered for keyness at five. The point of this parameter is to exclude words that will be identified as unusual simply because they happen not to have occurred in the 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; indeed, I would like to consider these in a further study. However, they are unlikely to be diagnostic of the character's general style, and so I shall not consider them here.[4] I selected the log-likelihood test for significance. (I repeated the analysis with the chi-square test: the same results were revealed with only minor and occasional differences in the ranking, which have no effect on my commentary below). I set the probability value at smaller than or equal to 0.05 (i.e. there is a 5% chance or less of the result being a fluke), a typical value for the social sciences.

5. Analysis and results: Keywords in *Romeo and Juliet*

A note on raw word frequencies

Before examining the keywords results, I shall briefly consider simple raw frequencies for each of the characters. The point of this is to justify why we need to engage in a more sophisticated analysis, such as keyword analysis. Table 2 displays the top ten rank-ordered word frequencies for the six characters. For comparative interest, I have included the ten highest word frequencies overall in the play, and also the ten highest word frequencies in present-day spoken English and present-day written English. Of course, comparing an Early Modern

English play-text with present-day general spoken language and general written language requires much caution!

Table 2. The top 10 rank-ordered word frequencies for six characters in *Romeo and Juliet* (Present-day data taken from Leech, G., Rayson, P. and Wilson, A. (2001))

Romeo	*Juliet*	*Capulet*	*Nurse*	*Mercutio*	*Friar L.*	*Overall in the play*	*Pres-day Spoken English*	*Pres-day written English*
And (136)	I (138)	To (61)	I (70)	A (85)	And (93)	And (734)	The	The
I (132)	To (113)	You (49)	A (61)	The (85)	The (83)	The (714)	I	Of
The (117)	And (104)	And (48)	And (61)	Of (57)	To (67)	I (589)	You	And
To (97)	My (92)	A (45)	The (56)	And (53)	In (51)	To (551)	And	A
My (85)	The (84)	My (45)	You (55)	To (36)	Thy (51)	A (473)	It	In
That (84)	That (82)	I (44)	To (45)	That (33)	Thou (46)	Of (389)	A	To (inf.)
A (78)	Thou (71)	Is (39)	It (39)	I (31)	Of (43)	My (361)	's	Is
Of (77)	Is (68)	The (37)	Is (34)	Is (31)	Is (37)	That (354)	to	To (prep.)
Me (73)	A (68)	Her (29)	My (33)	In (30)	That (36)	Is (342)	of	Was
In (72)	Be (59)	Not (29)	O (26)	Thou (27)	A (33)	In (321)	that	It

What will be clear from this table is that many words are common to many characters. Thus, they fail to discriminate between characters: they are not style markers. That said, there are some differences that are noteworthy. For example, a unique feature of the Nurse's list is the presence of the interjection 'o'. This would reflect the idea that the Nurse is a rather emotional character, and we will see more evidence of this later. Focusing on rank-order, it is interesting to note that Mercutio's top four words are identical to the words for present-day written English, and Friar Lawrence's top four words appear in the top six for present-day written English. In contrast, the other characters all have a first or second pronouns in the top four, as does the top four for present-day spoken English. It has been suggested that first and second pronouns are features of interaction (e.g. Biber 1988). Romeo, Juliet, Capulet and the Nurse are more interactive characters than Mercutio and Friar Lawrence, who both tend to extol forth regardless of other characters on stage. Also, the fact that Mercutio's word frequencies have some similarities with 'writtenness' is not surprising, since he has an elaborate rhetorical style, which we shall reveal later.

Of course, one can refine such frequency tables further. A common strategy, for example, is to remove grammatical or function words from the lists, so that only the frequencies of content words are shown. I was reluctant to follow this

path, because I am not convinced that grammatical words contain no evidence of style. Indeed, my keywords analysis will show the contrary.

Keywords

Table 3 displays all the keywords revealed for each of the six characters. Positive keywords are keywords that appear because they are unusually frequent; negative keywords are keywords that appear because they are unusually infrequent.[5]

The keywords facility in *WordSmith Tools* automatically separates positive keywords from negative, each rank-ordered according to keyness (i.e. how statistically unusual they are compared with the reference corpus). The fact that there are fewer negative keywords compared with positive keywords is not surprising: it is easier to do more than the norm established in a reference corpus than to less than that norm, particularly when the reference corpus is small. I do not have space to comment on each keyword. Instead, I shall comment on particularly salient or interesting trends for each character. I shall comment separately on the pronouns that appear as keywords in the following subsection. Not all keywords, however, relate to characterisation. Only by examining the usage of those keywords (i.e. by conducting qualitative analysis) can one determine whether a keyword has anything to do with characterisation. An important factor – though not necessarily a decisive one – in determining whether they relate to character or not is whether they are localised or well-dispersed throughout the play. Romeo's keyword 'banished', for example, only occurs 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. In the Table 3, raw frequencies of occurrence are supplied in round brackets (e.g. 'banished' occurs nine times). There is no simple correlation between these raw frequencies and whether a word is a positive or negative keyword, or how the items are ranked. Instead, raw frequencies may give some indication as to how well-dispersed the particular keywords are. For example, Mercutio's 85 instances of 'a', his most key (i.e. statistically unusual) keyword, are well-dispersed throughout the four scenes in which he speaks (i.e. I.iv, II.i, II.iv, III.i). In contrast, Mercutio's five instances of 'hare', his second most key keyword, all occur in a small cluster towards the end of Act II, scene iv. Of course, one cannot assume that a high frequency figure necessarily means an even dispersion throughout the play (or, more precisely, a particular character's talk), though it offers a good indication.

My discussion below is largely focused on keywords which I have established are well-dispersed in a particular character's talk.

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

	Romeo	*Juliet*	*Capulet*	*Nurse*	*Mercutio*	*Friar L.*
Positive keywords	Beauty (10) Blessed (5) Love (46) Eyes (14) More (26) Mine (14) Rich (7) Dear (13) Yonder (5) Farewell (11) Me (73) Sick (6) Lips (9) Stars (5) Fair (15) Thine (7) Hand (11) Banished (9)	If (31) Or (25) Sweet (16) Be (59) News (9) My (92) Night (27) I (138) Would (20) Yet (18) Thou (71) Words (5) Name (11) Nurse (20) Tybalt's (6) Send (7) Husband (7) That (82) Swear (5)	Go (24) Wife (10) Thank (5) Ha (5) You (49) T (5) Thursday (7) Her (29) Child (7) Welcome (5) We (15) Haste (6) Gentlemen (5) Tis (11) Our (13) Make (10) Now (15) Daughter (5) Well (13)	Day (22) He's (9) You (55) Quoth (5) Woeful (6) God (12) Warrant (7) Madam (10) Lord (11) Lady (16) Hie (5) It (39) Your (21) Faith (7) Said (6) Ay (90 She (21) About (5) Ever (5) Sir (13) Marry (7) Ah (6) Fall (5) Well (13)	A (85) Hare (5) Very (11) Of (57) He (20) The (85) O'er (5)	Thy (51) From (23) Thyself (5) Mantua (6) Part (7) Heaven (10) Forth (5) Her (30) Alone (6) Time (10) Married (7) Letter (5)
Negative keywords	You (14) Romeo (5) He (11) Go (7)	Her (5) The (84) You (27) And (104) Go (6)	Thou (7) That (13) The (37) Of (21) And (48)	With (12) Thou (11)	My (13) I (31) What (5)	I (32) You (16) A (33) Have (5) My (26)

Romeo's top three keywords 'beauty', 'blessed' and 'love' seem to match one's intuitions about this character: he is the lover of the play. Other keywords, such as 'dear', 'stars' and 'fair' fit his 'love talk' style. For example (all keywords are underlined in examples cited):

> She hath, and in that sparing makes huge waste; For beauty, starv'd with her severity, Cuts beauty off from all posterity. She is too fair, too wise, wisely too fair, To merit bliss by making me despair: She hath forsworn to love, and in that vow Do I live dead that live to tell it now. (I.i)

Keywords relating to body parts – 'eyes', 'lips' and 'hand' – underlie Romeo's concern with the physical:

> 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)

Interestingly, as is clear from the quotation above, Romeo often reflects on his own body parts. This hint at his egocentric nature is also reflected in the pronominal keywords, as we shall see in the following section.

Juliet's most 'key' keyword, 'if', is striking, because, unlike many of Romeo's keywords, it does not seem so obviously guessable, partly because it is a grammatical word. Here are some examples:

> If he be married, / My grave is like to be my 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' seems to reflect the fact that Juliet is in a state of anxiety for much of the play. Other keywords support this. 'Yet', another grammatical word, is a case in point:

> Tis almost morning; I would have thee gone; And yet no further than a wanton's bird […] (II.ii.) [whether Romeo should go]

> I fear it is: and yet, methinks, it should not, For he hath still been tried a holy man (IV.iii.) [whether the Friar has supplied sleeping potion or poison]

'If' and 'yet' create a syntactic style that is meaningful: it articulates Juliet's anxieties. This style is supported by other keywords, such as the subjunctive 'be'

(see the first instance in first example for Juliet above), and the modal 'would' (see the penultimate example for Juliet above, where it expresses her mixed wishes).

Capulet's most 'key' keyword is 'go'. This is usually an imperative command directed variously to Tybalt (I.v.82), Paris (III.iv.31), the Nurse (III.v.171), his servants (IV.11.2), Juliet (IV.ii.9), and Lady Capulet (IV.ii.41). There are no surprises here: as Capulet is the head of a noble household, his function in the play is largely to direct the others. The keywords 'make' and 'haste' are also part of this directive pattern, as can be seen from the following quotation:

> Go wake Juliet, go and trim her up; I'll go and chat with Paris. Hie, make haste, Make haste; the bridegroom is come already: Make haste, I say. (IV.iv)

Of course, it is Capulet's conspicuous failure to direct Juliet that constitutes part of the tragedy of the play.[6]

An interesting pattern in the Nurse's keywords is that many are what have been referred to as 'surge features', a term that refers to linguistic items which reflect 'outbursts of emotion' (Taavitsainen 1999). The Nurse's keywords which are clearly surge features include: 'god', 'warrant', 'faith', 'marry' and 'ah'. The following quotation illustrates the usage of three of these:

> Mistress! What, mistress! Juliet! Fast, I warrant her, she.
> Why, lamb! Why, lady! Fie, you slug-a-bed!
> Why, love, I say! madam! sweetheart! Why, bride!
> What, not a word? You take your pennyworths now.
> Sleep for a week; for the next night, I warrant,
> The County Paris hath set up his rest
> That you shall rest but little. God forgive me!
> Marry, and amen! How sound is she asleep! (IV.v.)

These surge features are not in fact indicators of transitory emotional reactions to circumstances. All of the surge features listed above occur in at least four scenes (the Nurse speaks in 11 scenes). The Nurse is dispositionally emotional. This is not to say, of course, that the context cannot trigger keywords which are symptoms of emotion. The Nurse's most key keyword is 'day' and her fifth most key keyword is 'woeful'. These are localised keywords: 'woeful' only occurs in Act IV scene v, and 15 of the 22 instances of 'day' occur in that scene. Here, the Nurse discovers Juliet apparently dead:

O wo, O woeful, woeful, woeful day,
Most lamentable day, most woeful day,
That ever, ever, I did yet behold.
O day, O day, O day, O hateful day,
Never was seen so black a day as this:
O woeful day, O woeful day. (IV.v.)

The Nurse, already an emotional character, reacts with extreme emotion (compare, for example, Capulet's rather more controlled and sophisticated rhetorical reaction). Finally, it is worth noting that many of the Nurse's keywords reflect her more colloquial register. This includes the surge features listed above, the item 'ay', the vocatives 'madam', 'lord', 'lady', and 'sir', and the speech report verbs 'quoth' and 'said' (which are evidence of the fact that the Nurse delights in oral narratives). Shakespeare is well-known for attributing a more colloquial style to characters of the lower social orders (see, for example, Gilbert 1979: chapter 2, for a discussion of 'high-style', 'middle-style' and 'low-style' in Shakespeare).

Mercutio's most key keyword is 'a', his fourth is 'of' and his sixth is 'the'. On the face of it, this appears to be an unexciting result. Note that such central grammatical items are often deleted when raw word frequency lists are considered, as I mentioned above. In fact, they reveal an important aspect of Mercutio style. They confirm an observation I made earlier that Mercutio has a more 'written' and less interactive style. More specifically, he has a 'noun-y' style. He has a tendency to use lists of noun phrases or prepositional phrases, as can be seen from the quotation below:

Ben. Why, what is Tybalt?
Mer. More than Prince of Cats. O,
He's the courageous captain of com-
pliments. He fights as you sing prick-
song: keeps time, distance, and propor-
tion; he rests his minim rests, one, two,
and the third in your bosom; 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.iii.)

Mercutio is in the play to give dazzling rhetorical displays, as well as to raise the emotional temperature and further the plot by getting killed. The dramatic focus is not on his social relations with other characters.

Friar Laurence is a man of the Church, and hence the presence of 'heaven' as a keyword is not surprising. However, his three most key keywords, 'thy', 'from' and 'thyself', are at first sight more puzzling. Consider this quotation:

> The sun not yet thy sighs from heaven clears,
> Thy old groans yet ring in my ancient ears;
> Lo! Here upon thy cheek the stain doth sit
> Of an old tear that is not washed off yet.
> If e'er thou wast thyself and these woes thine […] (II.iii)

Friar Laurence is not only the play's agony aunt, he is also an emotional mirror: he articulates the traumas Romeo and Juliet are suffering. I will comment on the fact that he uses 'thy' and 'thyself', as opposed to 'your' and 'yourself', in the following section. His other keywords, notably, 'Mantua', 'letter', relate to the role he plays in facilitating the plot.

Pronominal patterns in the keywords

In Table 3 above, pronouns were revealed as keywords, both positive and negative, for each character. Table 4 displays the rank-ordered pronominal keywords for each of the six characters.

Table 4. Rank-ordered pronominal patterns in the speech of six characters in *Romeo and Juliet*

	Juliet	*Romeo*	*Capulet*	*Nurse*	*Mercutio*	*Friar L.*
Positive keyword pronouns	My I Thou	Me Mine Thine	You We Tis Our	He's You It Your She	He	Thy Thyself Her
Negative Keyword pronouns	You	You He	Thou	thou	My I	I You My

Some interesting trends are apparent in these pronominal keywords. Table 5 summarises these trends.

Table 5. The pronominal preferences of six characters in *Romeo and Juliet* (round brackets indicate secondary preferences)

	Juliet	*Romeo*	*Capulet*	*Nurse*	*Mercutio*	*Friar L.*
Person	1st (2nd)	1st (2nd)	2nd (1st)	3rd (2nd)	3rd	2nd (3rd)
1st person singular / plural	S	S	P	-	-	-
Thou / you	T	T	Y	Y	-	T

For both Romeo and Juliet, the top two pronominal keywords are first person singular, whilst the pronoun ranked third is second person. It is no surprise that Romeo and Juliet use first and second pronouns: they are at the heart of the social interaction in the play. Interestingly, the subjective first person pronoun 'I' appears in Juliet's list, but not in Romeo's, where instead we find the objective first person pronoun 'me'. Although this is not conclusive evidence, it is consistent with the idea that Juliet spends much time in the play bearing her soul (cf. instances in the examples above for Juliet), whereas Romeo is much more conscious of his own role as a lover and of the effect of circumstances upon him. Consider the following examples:

> O, wilt thou leave me so unsatisfied? (II.ii.)
>
> […] O sweet Juliet!
> Thy beauty hath made me effeminate,
> And in my temper soft'ned valour's steel! (III.i.)
>
> Thou canst not speak of that thou dost not feel:
> Wert thou as young as I, Juliet thy love,
> An hour but married, Tybalt murdered,
> Doting like me, and like me banished […]. (III.iii.)
>
> In faith, I will. Let me peruse this face:
> Mercutio's kinsman, noble County Paris!
> What said my man when my betossed soul
> Did not attend him as we rode? I think
> He told me Paris should have married Juliet:
> Said he not so? or did I dream it so?
> Or am I mad, hearing him talk of Juliet,
> To think it was so? O! give me thy hand,
> One writ with me in sour misfortune's book:
> I'll bury thee in a triumphant grave [...] (V.iii.)

Capulet also uses first and second person pronouns, but importantly in reverse order of preference – 'you' is most key. This supports the idea that Capulet's major role in the play is to direct other people. When Capulet does use first person pronouns, note that they are plural (i.e. 'we', 'our'). This relates to the fact that Capulet, the head of the household, often speaks on behalf of other people, as well as himself. In the following instance of 'we', he refers to Montague and himself:

> But Montague is bound as well as I,
> In penalty alike; and 'tis not hard, I think,
> For men so old as we to keep the peace. (I.ii)

In the first and second instances of the plural first person pronouns below, Capulet is probably referring to Lady Capulet and himself as they both encounter Paris, whilst the third refers to all people:

> Things have fall'n out, sir, so unluckily,
> That we have had no time to move our daughter:
> Look you, she lov'd her kinsman Tybalt dearly,
> And so did I: well, we were born to die. (III.iv)

In this example the reference of the plural first person pronouns is unclear:

> All things that we ordained festival,
> Turn from their office to black funeral;
> Our instruments to melancholy bells,
> Our wedding cheer to a sad burial feast,
> Our solemn hymns to sullen dirges change,
> Our bridal flowers serve for a buried corse,
> And all things change them to the contrary. (IV.v.)

The pronouns could refer to: (1) Capulet alone (what we would call today the 'royal we'), (2) Lady Capulet and himself, or (3) the Capulet household. What is clear is that Capulet steadfastly avoids pronouns signalling individual reference to himself alone.

Third person pronouns appear as most key for the Nurse and Mercutio, and of secondary keyness for Friar Laurence. (Note that 'he' appears as a negative keyword for Romeo: he avoids this third person pronoun). This reflects the fact that these characters tend to discourse about third parties, rather than with second parties. This is clear in the quotation given for Mercutio above, and in such discourse as this from the Nurse, as she recounts Juliet's youth:

> […] Yet I cannot choose but laugh,
> To think it should leave crying, and say 'Ay.'
> And yet, I warrant, it had upon its brow
> A bump as big as a young cockerel's stone;
> A parlous knock; and it cried bitterly:
> 'Yea,' quoth my husband, 'fall'st upon thy face?
> Thou wilt fall backward when thou com'st to age;
> Wilt thou not, Jule?' it stinted and said 'Ay.' (I.iii.)

(The Nurse's strongly key keyword 'he's' is localised: it is a reaction to Romeo's death, e.g. 'he's dead, he's dead').

The fact that Friar Laurence has a second person pronoun as his most key pronominal keyword merits comment. But note that he only has the possessive form of the second person pronoun (or the possessive reflexive form) as positive keywords. In fact, 'you' appears as a negative keyword. Friar Laurence's role is not to interact with the others, but to describe them, as can be seen from the quotation in the previous section. Also, we might note that for both Friar Laurence and Mercutio 'I' and 'my' appear as negative keywords. These are the most minor characters of the six. As I have been arguing above, they are not in the play to volunteer information about themselves, but to describe others and facilitate the plot.

Finally, I shall briefly comment on the second person pronominal variants. Elizabethan English offered a choice between two forms for the second person pronouns: the plural forms 'ye', 'you', 'your', 'yours' and 'yourself', and the singular forms 'thou', 'thee', 'thy', 'thine' and 'thyself' (hereafter referred to collectively as you-forms and thou-forms respectively). The variant chosen could have significant social or pragmatic implications. The usage of these variants is a matter of great controversy – a controversy in which I cannot engage here. What we can note from the keywords analysis is that certain characters prefer you-forms and avoid thou-forms, whilst others do the opposite. Romeo and Juliet prefer thou-forms. Contrary to this, Brown and Gilman (1960) predict that high status social equals use you-forms. However, other researchers consider you-forms as dispassionate and emotionally unmarked, and thou-forms as expressive of negative emotions (e.g. anger and contempt) or positive emotions (e.g. affection and love) (e.g. McIntosh 1963, and Mulholland 1987). This is consistent with the idea that the thou-forms favoured by Romeo and Juliet mark the discourse of intimacy, or, more specifically here, love talk. This would also explain why Friar Laurence has a preference for thou-forms, since he engages in intimate and emotionally charged discourse. Both Capulet and the

Nurse prefer you-forms. Capulet's usage would fit Brown and Gilman's (1960) prediction that social superiors use you-forms. Regarding the Nurse, Brown and Gilman (1960) predict that low status individuals use you-forms to social superiors. The Nurse mostly interacts with people of much higher social status (e.g. Juliet, Romeo, Lady Capulet): hence her preference for you-forms. It may be objected that the Nurse is an emotional character, as I have already observed, and consequently this should, if the above arguments are correct, weight the Nurse's usage in favour of thou-forms. However, when the Nurse is at her most emotional (e.g. in Act IV, scene v, when she finds Juliet apparently dead), she tends not to address anybody but to give personal vent to her own emotions.

6. Conclusion

I started this paper by discussing the notions of 'style markers', 'style reminders' and 'keywords'. I argued that the notion of style markers is synonymous with that of keywords, but that the term keywords links in with corpus linguistics research in general and the computer program *Wordsmith Tools* (Scott 1999) in particular. One specific innovation of my work is that I have analysed dialogue for keywords. This was made possible by adopting a relatively simple tagging system. The main aim of this paper was to show how a keywords analysis offers an empirical way of establishing lexical and grammatical character patterns. Specifically, I aimed to reveal lexical and grammatical character patterns in Shakespeare's *Romeo and Juliet*. In some cases, my analysis provided solid evidence for what one might have guessed (e.g. Romeo's keywords 'beauty' and 'love'); in others, it revealed what I think would be very difficult to guess but fits well a possible interpretation (e.g. Juliet's keywords 'if' and 'yet'). I also demonstrated how keywords analysis also offers a way into analysing function words, such as pronouns, and accounting for their contribution to style and meaning.

Of course, as with any study, there are a number of limitations that need to be borne in mind (in no particular order):

- I did not attempt to lemmatize the word forms in my data, so that, for example, 'loves' would form part of the word count of 'love'. I did not do this because: (a) lemmatization programs tend to be inaccurate, and (b) lemmatisation can obscure shades of differences in meaning.

- Contractions would have been counted separately, so that, for example, the first person singular pronoun in the contraction 'I'll' would not have been counted along with individual instances of the pronoun.
- Keywords involve statistical deviation from a relative norm. This is only one kind of deviation. Levin (1963), for example, distinguishes between 'determinate' and 'statistical' deviation, or, similarly, Leech (1985) distinguishes between an 'absolute norm' and 'relative norm'. For example, an interruption can break an absolute norm (a social rule); similarly, spelling an English word 'ndfjh' breaks an absolute norm (a structural rule of English). Of course, there is an intimate relationship between these two types of norms: deviations from absolute norms tend to occur relatively infrequently. The point here is that a keyword analysis focuses squarely on statistical deviations from a relative norm, and ignores the significance of relatively infrequent deviations from absolute norms.
- Related to the above, keywords analysis does not consider hapax legomena – one-off occurrences of words.

The possibilities for future research are numerous. One area is to consider is relationships between keywords. Scott (1999) has developed, for example, the notion of 'key-keywords' (keywords 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). I was not able to pursue these notions in my data, on account of the small number of words in each character file, but this clearly would be an avenue to pursue with larger quantities of data. Another obvious area for development is to apply keywords analysis to other kinds of data, such as particular registers, dialects, media, documents or writings.

References

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

Blake, N.F. (1983) *Shakespeare's Language: An Introduction*, London: MacMillan.

Brown, P. and Levinson, S.C. (1987) *Politeness: Some Universals in Language Usage*, Cambridge: Cambridge University Press.

Brown, R. and Gilman, A. (1960) 'The pronouns of power and solidarity', in Sebeok, T.A. (ed.), *Style in Language*, Cambridge, Mass.: MIT Press, 253-76.

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.

Enkvist, N.E. (1964) 'On defining style', in Enkvist, N.E., Spencer, J. and Gregory, M. (eds.) *Linguistics and Style*, Oxford: Oxford University Press, 1-56.

Enkvist, N.E. (1973) *Linguistic Stylistics*, Berlin: Mouton.

Gilbert, A.J. (1979) *Literary Language from Chaucer to Johnson*, London: MacMillan.

Leech, G.N. (1981) *Semantics*, Middlesex: Penguin Books.

Leech, G., Rayson, P. and Wilson, A. (2001) *Word Frequencies in Written and Spoken English Based on the British National Corpus.*

McIntosh, A. (1963) '"As You Like It": A grammatical clue to character', *A Review of English Literature*, 4, 2, 68-81.

Mulholland, J. [1967] (1987) '"Thou" and "you" in Shakespeare: A study in the second person pronoun', reprinted in Salmon, V. and Burness, E. (eds), 153-161.

Page, N. (2nd edn.) (1988) *Speech in the English Novel*, Houndmills, Hampshire: MacMillan.

Salmon, V. and Burness, E. (eds) (1987) *A Reader in the Language of Shakespearean Drama*, Amsterdam and Philadelphia: John Benjamins.

Scherer, K.R. (1979) 'Personality markers in speech', in Scherer, K.R. and Giles, H. (eds), *Social Markers in Speech*, Cambridge: Cambridge University Press, 147-209.

Scott, M. R. (1999) *WordSmith Tools*, Oxford: Oxford University Press (see also http://www.liv.ac.uk/~ms2928/).

Taavitsainen, I. (1999) 'Personality and styles of affect in *The Canterbury Tales*', in Lester, G. (ed.), *Chaucer in Perspective: Middle English Essays in Honour of Norman* Blake, Sheffield: Sheffield Academic Press, 218-234.

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

[1] The culmination of this work is Culpeper (2001).

[2] The switch-on tag can be introduced automatically by doing a 'find and replace' operation on the original speaker identification in the text (e.g. FIND Sam. + REPLACE WITH <SAM>). The switch-off tags are a little more problematic. I introduced them manually. However, it should be possible to devise a program to introduce them automatically, since their position can be identified by the fact that they precede the switch-on tag for a different speaker (this does not apply where stage directions appear, so some way of identifying stage directions would be required).

[3] To clarify, each of the six characters was compared with the other characters in the play, excluding the particular character in question (thus, Romeo would be compared with a data set consisting of the other characters but excluding Romeo himself). This was so that the comparative data would be an independent variable.

[4] In studies I have come across in corpus linguistics, 'ten' seems to be a favourite minimum frequency. I experimented with various settings. My character data sets are relatively small, and so 'five' seemed to work best.

[5] The keywords facility in *WordSmith Tools* automatically separates positive keywords from negative, each rank-ordered according to keyness (i.e. how statistically unusual they are compared with the reference corpus). The raw frequencies of occurrence are supplied in round brackets. There is no simple correlation between these raw frequencies and whether a word is a positive or negative keyword, or how the items are ranked. Instead, raw frequencies may give some indication as to how 'general' the particular keywords are. However, one cannot assume an even dispersion throughout the play. I will draw attention to the importance of dispersion in my discussion.

[6] Readers may wonder what Capulet's keyword 't' is supposed to represent. In fact, it is a reduced form of 'to'. Not much significance can be attributed to the fact that Capulet has this graphological oddity. Also, we should remember that this reduced form would almost certainly have been put there by a compositor or editor, not by Shakespeare.