ÉTUDES DE STYLISTIQUE ANGLAISE

Études de stylistique anglaise

4 | 2013
Style in Fiction Today

Virginia Woolf meets Wmatrix

Geoffrey Leech

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/esa/1405
DOI: 10.4000/esa.1405
ISSN: 2650-2623

Publisher
Société de stylistique anglaise

Printed version
Date of publication: 1 March 2013
Number of pages: 15-26
ISSN: 2116-1747

Electronic reference
Geoffrey Leech, « Virginia Woolf meets Wmatrix », *Études de stylistique anglaise* [Online], 4 | 2013, Online since 19 February 2019, connection on 02 May 2019. URL : http://journals.openedition.org/esa/1405 ; DOI : 10.4000/esa.1405

VIRGINIA WOOLF MEETS WMATRIX[1]

Geoffrey Leech
Lancaster University, U.K.

Résumé : Le logiciel WMatrix, créé par Paul Rayson, permet une analyse stylistique comparée d'un texte au regard d'un corpus de référence, c'est-à-dire un corpus représentant un « style d'anglais » pertinent pour la comparaison. Pour cette étude expérimentale, j'ai choisi la nouvelle de Virginia Woolf intitulée « The Mark on the Wall » (1917) comme texte soumis à l'étude. Cette étude s'est révélée assez concluante en ce qu'elle a permis de mettre en lumière des mots-clés ainsi que d'autres items que j'avais, de manière impressionniste, jugés pertinent d'un point de vue stylistique et thématique.

Mots-clés: WMatrix – corpus – analyse stylistique

Stylistic analysis is essentially a comparative process. An automatic method of comparing bodies of text in order to characterize their 'differentness' is provided by the Wmatrix software developed by Paul Rayson (for details, see Rayson 2008; also http://ucrel.lancs.ac.uk/wmatrix/). For my purposes, as I am interested here in the stylistic analysis of a single text, the comparison will be between that single text (the **focal text**) and a corpus (**the reference corpus**). The question is: How far can this automated procedure help to identify salient features of literary style? How far can phenomena which are statistically salient

1 This article, although entirely written by me about research I undertook, was largely written as part of another paper which has been awaiting publication for three years: Geoffrey Leech, Nicholas Smith and Paul Rayson (forthcoming) 'English style on the move: changing stylistic norms in the twentieth century'. In Merja Kytö (ed.) *English Corpus Linguistics: Crossing Paths.* Amsterdam: Rodopi. I am grateful to my collaborators Paul Rayson and Nick Smith for their help, especially on the use of WMatrix.

in the text be considered foregrounded from the point of view of literary theme and appreciation?

1. Virginia Woolf's 'The Mark on the Wall': Our focal text

'The Mark on the Wall', written in 1917, might be described as a story in which nothing happens – where nothing happens, that is, except in the mind of the narrator. (We use the term 'narrator' here, although it is the *inner* voice of the narrator that we experience throughout the story.) The narrator, sitting down after tea, notices a mark on the wall. Her mind explores in a myriad ways the significance of that mark – what it might be, and where it came from. This train of thought leads her by digressions of memory and imagination to such topics as the preceding occupants of the house – the nature of life – life after death – the oddities of experience – the mysteries of existence – always following the stream of the narrator's consciousness. Every so often, however, the narrator's attention comes back to the mark on the wall – and at last, she learns what it is. To give the flavour of the text, here are its opening paragraph and the final few lines:

> Opening paragraph:
>
> Perhaps it was the middle of January in the present year that I first looked up and saw the mark on the wall. In order to fix a date it is necessary to remember what one saw. So now I think of the fire; the steady film of yellow light upon the page of my book; the three chrysanthemums in the round glass bowl on the mantelpiece. Yes, it must have been the winter time, and we had just finished our tea, for I remember that I was smoking a cigarette when I looked up and saw the mark on the wall for the first time. I looked up through the smoke of my cigarette and my eye lodged for a moment upon the burning coals, and that old fancy of the crimson flag flapping from the castle tower came into my mind, and I thought of the cavalcade of red knights riding up the side of the black rock. Rather to my relief the sight of the mark interrupted the fancy, for it is an old fancy, an automatic fancy, made as a child perhaps. The mark was a small round mark, black upon the white wall, about six or seven inches above the mantelpiece.
>
> Ending:
>
> ... – but something is getting in the way ... Where was I? What has it all been about? A tree? A river? The Downs? Whitaker's Almanack? The fields of asphodel? I can't remember a thing. Everything's moving, falling, slipping, vanishing ... There is a vast upheaval of matter. Someone is standing over me and saying:
>
> 'I'm going out to buy a newspaper.'
> 'Yes?'
> 'Though it's no good buying newspapers. Nothing ever happens. Curse this war; God damn this war! ... All the same, I don't see why we should have a snail on our wall.'
> Ah, the mark on the wall! It was a snail.

2. Comparing the focal text and a reference corpus

The focal text, 'The Mark on the Wall', will be compared quantitatively with a reference corpus which should be representative to some degree of the variety from which the text is taken. However, there are obviously different degrees of generality in defining the language variety meant to act as a reference standard. We have decided to use three different 'reference varieties' (the choice being determined, obviously, by the availability of suitable texts in electronic form):

(A) a rather specific variety, resembling the focal text in three ways: it consists of (1) fiction writing (2) by women writers (3) published in 1917. On the other hand, this reference corpus is limited in representativeness, as it contains only three novels, the work of three authors.[2]

(B) A more general corpus of fiction, consisting of category K (General Fiction) in the Fiction subcorpus of the B-LOB corpus (a member of the Brown Family of corpora representing written (printed) British English over the period 1928-1934). This is more widely representative than (A), as it contains 29 text samples by different authors. However, it is less closely matched than (A) in time of publication, as the samples date from 1928-34.

(C) A very general corpus, sampled from the written (published) English of roughly the same period and national variety (British English of the beginning of the twentieth century) as the focal text. For this we used a third of the as yet incomplete 1901±3 corpus of the Brown family, covering all four of the subcorpora Press, General Prose, Learned and Fiction.[3] The corpus is not closely matched with 'The Mark on the Wall' temporally – indeed it is a worse match than (B), but may be considered more broadly representative than the other two of the written prose of the period, containing 166 text samples across a wide range of fiction and non-fiction writing.[4]

---

2 A selection of notable novels published in the same year as 'The Mark on the Wall' are listed at 'Literature in 1917', Wikipedia. The following three were found to be available from Project Gutenberg and other on-line resources: Florence Barclay, *The White Ladies of Worcester*; Mrs Humphrey Ward, *Missing*; Edith Wharton, *Summer*. Two of the authors are British and one (Wharton) American.

3 The one-third 1901 corpus contained one-third of each subcorpus, and each text category in proportion to their representation in the Brown-family corpus when complete. Within each text category, the texts were also matched in topic and publication with the corresponding parts of B-LOB, LOB and F-LOB.

4 In terms of Wmatrix word counts, the size of the focal text is 2,985 words, and the sizes of the reference corpora are: Three 1917 Novels: 269,842; 1901 Corpus: 342,448; B-LOB General Fiction: 56,703. Wmatrix word counts are generally slightly lower than other corpus tools because semantically meaningful chunks, e.g. idiomatic expressions, names, places, and phrasal verbs, are counted as one item.

In practice, none of our reference corpora are ideal; and one of the interests of this study was to discover how far the differences between the three reference corpora of increasing generality would produce different results.[5] So, what is the method of comparison?

The methodology employed by Wmatrix is broadly definable as an extraction from the data of **keywords**, or rather **key features**: that is, words or other features of the text which stand out or deviate, in a statistical sense, from the frequencies of the reference corpus. The statistical concept of **keywords** has become familiar in corpus linguistics since it was built into the popular corpus software package WordSmith Tools (Scott 2004), and has since been the basis of a considerable body of published research.[6] In the case of Wmatrix, however, this method has been extended further to grammatical word classes (parts of speech) and to semantic domains, as will be shortly explained. In other words, the comparison is not purely lexical.

To begin with keywords: by 'keyness' here is meant the words which are most distinctive of that text, as contrasted with the reference corpus. Keyness so understood is of variable strength, so that the output of this process of keyword extraction is a list, in which words are listed in order of keyness. Similar lists can be obtained for any other features of language automatically identifiable in the textual data. The general set of procedures involved in a research project of this kind can be listed as the four stages below:[7]

1. *Building the data*: corpus design and compilation (in the case of our Wmatrix investigation, this has already been sufficiently described in terms of our focal text and the three reference corpora).
2. *Annotating the data*: analysing the corpus linguistically, using particular annotation tools: in the case of Wmatrix, the two annotation tools used are

(a) the CLAWS part-of-speech (POS) tagger, and
(b) the USAS semantic domain tagger.

---

5 In Leech (2008: 168-76) two widely differing reference corpora were used – (a) three novels of the 1890s and (b) the General Fiction text category (K) of the B-LOB Corpus, dating from 1928-34. In view of their disparity, it was surprising that the overall analysis was closely similar for both corpora.

6 See the list of publications on Mike Scott's webpage http://www.lexically.net/publications/publications.htm

7 This is a simplified version of the five-stage process presented in Rayson (2008: 521).

Details of these tools are to be found on the UCREL (Lancaster) website at: http://ucrel.lancs.ac.uk/claws/ and http://ucrel.lancs.ac.uk/usas/.)[8]

3. *Retrieving*: extracting from the text data some analytic results, which may be displayed in a variety of formats for inspection or further processing. In the Wmatrix analysis, we are interested in three more or less standard listing formats:

> (a) concordances, which list the occurrences of a particular word (or other feature) in their contexts of occurrence,
> (b) frequency lists, which list words (or other features) in order of their frequency in a particular body of text data, and
> (c) keyness lists, which list words (or other features) in order of their keyness in a given textual comparison.

4. *Interpreting*: This is the only stage of the process which is essentially non-automatic ('manual'), although it can be aided by automatic procedures such as using the 'Sort' and 'Collocation' facilities of corpus software. Whereas stages 3(a) and 3(b) above are quantitative, stage 4 is qualitative: it makes use of the human ability to interpret texts and to explain the phenomena observed in them. In the case of the Wmatrix investigation, we may be interested here in examining the textual material more carefully, using especially the concordance displays, in order to explain the stylistic phenomena observed in the analysis.

We now have to focus on the third, 'Retrieving' stage above, in order to explain in a little more detail what the software does. At the same time, we will avoid going into technical detail, which can be studied in Rayson (2008) and on the UCREL webpages already cited.

To take the most basic case, the list of keywords is arrived at as follows:

> i) Two word frequency lists are compiled: a list for the focal text ('List X'), and a list for the reference corpus ('List Y').
>
> ii) List X and List Y are compared. This means that each word in List X is measured in terms of *comparative frequency* with the same word in List Y.[9] 'Comparative frequency' means that the raw count of a word's frequency is adjusted to a standard measure relative to corpus size, which in Wmatrix is the number of occurrences of the word as a percentage of all occurrences of words in the text/corpus.
>
> iii) Each word's keyness in the focal text is measured by a statistical formula, which calculates the degree to which the word is either 'over-represented' or 'under-represented' in

---

8 Note that these tools do not produce error-free output. The accuracy of CLAWS is in the region of 96-7%, and that of USAS is c. 91%. These accuracy rates, however, are high enough to provide a sound basis for key feature extractions, given that the most salient results show high statistical significance (see below).

9 The keyword list can include words which have 0 occurrences in List X or List Y. Negative keywords are normally less noticeable and interesting, but can be important – e.g. it is significant that 'The Mark on the Wall' makes very little use of third person pronouns such as *she* and *they*.

this text, as measured against the reference corpus. The normal understanding of keyness is that the word is *over*-represented, that is, is relatively more frequent in the focal text than in the reference corpus, to a certain high degree of statistical significance.[10]

iv) The words in List X are re-ordered in order of keyness. This means that the words at the top of the list are most distinctive of that text.

Concordance, frequency and key-feature lists of POS tags and semantic tags are extracted in the same way as the word lists described in 3(a)-(c) above. There are no particular difficulties in this, as the annotation (tagging) has meant that each word in each text is accompanied by label giving its grammatical and semantic classification.

3. Results: keywords, key POS tags, and key semantic domain tags

To begin with, Table 3 shows the top 12 keywords, in order, when 'The Mark on the Wall' is compared with each of the reference corpora.

Table 3 **Keywords: Words of abnormally high frequency in 'The Mark on the Wall'**

A. compared with three 1917 novels by women writers		B. compared with 1931 general fiction (category K of B-LOB)		C. compared with the ⅓ 1901±3 Brown-family corpus	
1. mark	7. worshipping	1. mark	7. of	1. mark	7. one
2. is	8. thoughts	2. is	8. nail	2. wall	8. I
3. one	9. of	3. wall	9. reality	3. Whitaker	9. Precedency
4. Whitaker	10. tree	4. thoughts	10. tablecloths	4. thoughts	10. mantelpiece
5. wall	11. Precedency	5. Whitaker	11. worshipping	5. tablecloths	11. nail
6. tablecloths	12. chancellor	6. one	12. tree	6. worshipping	12. tree

NOTE: Double underlining marks the words which are in the top 12 for all three comparisons. Single underlining marks the words which are in the top 12 for two of the three comparisons.

Perhaps the most striking result is the amount of agreement that the three reference corpora show, in spite of their very different composition. Comparisons with A and B share all of their top 10 key words (out of 12); A and C share 9 of the 12; and B and C share 11. Perhaps this is a mild reflection of the degree of generality of the corpora. It seems that the keyword methodology is robust in showing up the 'differentness' of a text without respect to the exact make-up of the reference corpus.

It is not surprising that *mark* is the 'keyest' of the keywords: it represents the theme of the story, as to a lesser extent does *wall*. These are words that, as

10 The significance measure used in Wmatrix is log likelihood, which is considered preferable to the more familiar chi-square test, and which is explained in Rayson (2008: 527-8) and at http://ucrel.lancs.ac.uk/llwizard.html .

we might imagine, occur relatively rarely in the reference corpora, and therefore their repeated use in 'The Mark' is salient, both statistically and thematically. Of the other words which occur in all three comparisons, *one* (typically used in the generic human sense) is perhaps a personal stylistic favourite of Virginia Woolf, representing as it does the objectification of the narrator's personal experiences, as illustrated in the following passage:

> because *one* will never see them again, never know what happened next ... as *one* is torn from the old lady about to pour out tea and the young man about to hit the tennis ball in the back garden of the suburban villa as *one* rushes past in the train.

We will not dwell on the items in this list, some of them uncommon words, like *Precedency*, which gain idiosyncratic prominence in Woolf's narrative – see Leech (2008: 168-71) for further discussion. But there are some interesting points to observe about the similarities and differences between the lists. For example, *is* is very much overrepresented when compared with the fictional reference corpora (but not with the more general reference corpus C), and this is probably because Woolf, in capturing the immediacy of the interior monologue, tells much of her story in the historic present, instead of using the past tense narrative convention of the majority of fictional writers. This choice of the present tense is understandably not so salient when compared with the full range of written texts (scientific, journalistic, etc.) in the 1901±3 corpus. On the other hand, the pronoun *I*, frequent in Woolf's first-person narrative, stands out as over-represented when compared with the cross-section of written texts in 1901±3, but is less salient in the two fiction reference corpora, where first person reference occurs frequently, for example in dialogue.

We move on now to the lists of key part-of-speech tags, reflecting the different grammatical choices made by Virginia Woolf as compared with the writers in the other reference corpora.

Table 4: **The most 'key' parts of speech in 'The Mark on the Wall'**

compared with three 1917 novels		compared with 1931 general fiction		compared with 1901 Brown family corpus (⅓)	
1. VVZ	7. DDQ	1. VVZ	7. VV0	1. PN1	7. NN2
2. NN2	8. PPIS1	2. NN2	8. AT	2. PPIS1	8. PNX1
3. PN1	9. PNX1	3. PN1	9. PNX1	3. VVZ	9. RGQ
4. VBZ	10. NPD1	4. VBZ	10. RPK	4. VVG	10. DDQ
5. IO	11. RPK	5. IO	11. RGQ	5. RPK	11. AT1
6. AT1	12. RGQ	6. DDQ	12. NPD1	6. VV0	12 PPH1

NOTE: As in Table 3, double underlining marks the tags which are in the top 12 for all three comparisons. Single underlining marks the tags which are in the top 12 for two of the three comparisons.

Key: AT – article neutral for number; chiefly the definite article *the*.
AT1 – singular article; chiefly the indefinite article *a/an*
DDQ – *wh*-determiner or *wh*-pronoun (e.g. *what, which*)
IO – the preposition *of*
NN2 – plural common noun (e.g. *tables, women, thoughts*)
NPD1 – singular weekday noun (e.g. *Sunday, Monday*)
PN1 – singular indefinite pronouns (e.g. *one, anything, nobody*)
PNX1 – indefinite reflexive pronoun (i.e. *oneself*)
PPH1 – third person personal pronoun *it*
PPIS1 – the first person subject pronoun *I*
RGQ – *wh*-adverb of degree (*how* when modifying another word)
RPK – *about* used in the expression *be about to*.
VBZ – present tense *–s* form of the verb *to be* (i.e. *is*)
VVG – *ing*-form of lexical verb (e.g. *saying, wishing*)
VVZ – present tense lexical verb ending in *–s* (e.g. *says, wishes*)
VV0 – present tense lexical verb not ending in *–s* (e.g. *say, find*)

The amount of shared 'key tags' between the comparisons here is the same: nine tags are shared by the top twelve in A, B and C. What brings A and B closer together, however, is the fact that the top four tags are the same and in the same order. As mentioned above, the present tense (represented in the keyness of the *s*-form of lexical verbs VVZ as well as of VBZ and VV0), is a distinctive feature of 'The Mark', as opposed to fiction written in the more conventional past-tense narrative. More difficult to explain is the second-keyest tag, the plural noun tag NN2; however, the following passage illustrates how Woolf's style may favour plural nouns in describing the multitudinous particularity of her experiential world:

> let me just count over a few of the *things* lost in one lifetime, beginning, for that seems always the most mysterious of *losses* – what cat would gnaw, what rat would nibble – three pale blue *canisters* of book-binding *tools*? Then there were the bird *cages*, the iron *hoops*, the steel *skates*, the Queen Anne coal-scuttle, the bagatelle board, the hand organ – all gone, and *jewels*, too. *Opals* and *emeralds*, they lie about the *roots* of *turnips*.

It is striking, also, that this passage contains four examples of another key tag, IO (representing the preposition *of* in the tagging system). The word, of course, has many functions – but its main function, in the most general terms, is to signal the interconnectedness of things. It is noticeable in this list that IO stands out as a key tag in relation to the fictional reference corpora A and B, but not in relation to the most general reference corpus C, which is predominantly non-fictional. Elaboration of noun phrases by means of *of* is likely to be a characteristic of informational texts, which oddly here seem to be more akin to Woolf's own elaborative style. Of the other key tags, we will comment only on PN1, PNX1 and RGQ. PN1 chiefly represents the pronoun *one* already noted as favoured in 'The Mark'; and PNX1, normally a very rare tag (representing the word *oneself*) stands out in this text even though there are

only two occurrences of it. RGQ represents the adverb *How* as a modifier, in this text especially associated with exclamations:

> How readily our thoughts swarm...
> How shocking, and yet how wonderful it was to discover...
> How peaceful it is down here.

This construction may, indeed be another authorial favourite of Virginia Woolf, indicative of the narrator's (or a character's) characteristic emotional involvement in her subject matter.[11]

The third level of analysis, that of semantic tagging, produces lists of key semantic domains as follows:

Table 5: **The most 'key' semantic domains in 'The Mark on the Wall'**

compared with three 1917 novels	compared with 1931 general fiction	compared with 1901 Brown-family corpus (⅓)
1. General & abstract (*thing, things*) 2. Evaluation: authentic (*real, reality, really* 3. Plants (*tree, roots, stalk, flower*) 4. Life and living things (*life, lives*) 5. Colours & colour patterns (*blue, light, colour*) 6. Mental object; conceptual (*thought, thoughts, ideas*) 7. Smoking and non-medical drugs (*cigarette(s)*) 8. Living creatures: animals, birds (*cat, snail*) 9. Solid materials (*coals, glass, iron, emeralds*) 10. No kin (*illegitimate*) 11. Comparing (*compare, comparison* 12. Probability (*perhaps*)	1. Evaluation: authentic (*real, reality, really*) 2. Plants (*tree, roots, stalk, flower*) 3. Solid materials (*coals, glass, iron, emeralds*) 4. Colours & colour patterns (*blue, light, colour*) 5. General appearance & physical properties (*mark*) 6. General & abstract (*thing, things*) 7. Mental object; conceptual (*thought, thoughts, ideas*) 8. Living creatures: animals, birds (*cat, snail*) 9. Objects generally (*bowl, rock, hoops*) 10. Strong obligation & necessity (*must, should*) 11. Smoking and non-medical drugs (*smoke(s), cigarette(s)*) 12. Furniture and household fittings (*chair, table*)	1. General & abstract (*thing, things*) 2. Colours & colour patterns (*blue, light, colour*) 3. Evaluation: authentic (*real, reality, really*) 4. Plants (*tree, roots, stalk, flower*) 5. Life and living things (*life, lives*) 6. Parts of buildings (*wall, room, door*) 7. Furniture and household fittings (*chair, table*) 8. Smoking and non-medical drugs (*cigarette(s)*) 9. Thought, belief (*think, believe, imagine*) 10. The universe (*world, moon*) 11. Like (*like(s), adoring, fancy*) 12. Living creatures: animals, birds (*cat, snail*)

NOTE: Here we use double- and single-underlining in the same way as for the preceding two tables, but we underline only the number showing a semantic tag's position in the Table.

11 It is worth mentioning that this exclamatory construction is associated with female speech, being used by more female speakers than male speakers in each age group in the conversational part of the British National Corpus.

Key semantic domains tell us something about the 'aboutness' of texts, rather than about their stylistic characteristics in the strict sense. They are therefore less relevant to style, and there is less agreement between the different reference corpus comparisons: only half of the key semantic domains listed are shared by all three lists. On the other hand, there are some features which are salient not so much in style as in the authorial world view. The domain of colour is high on the list of key domains in all three comparisons, as are the domains relating to the natural world: 'Plants' and 'Living creatures'. Readers of Virginia Woolf will probably agree that these traits have a 'key' role in her writing. Other, more abstract domains are more difficult to interpret, but arguably reflect her exploration of the nature of reality and the ontological concerns of her writing. At the other extreme, the domain of 'Smoking' must be regarded as incidental to the text, in that it results from the semantic tagging of four words only: one of the drawbacks of choosing such a short focal text for analysis is that such haphazard results can occur. Here is another excerpt, which contains a reference to smoking, but is also relevant to some other key features:

> Even so, life isn't done with: there are a million patient watchful lives for a tree, all over the world, in bedrooms, in ships, on the pavement, lining rooms, where men and women sit after tea, smoking cigarettes. It is full of peaceful thoughts, happy thoughts, this tree.

This passage illustrates representation of some of the key features high on the list above: Plants (*tree*), Life and living things (*life, lives*), Mental object; conceptual (*thoughts*), Parts of buildings (*bedrooms, rooms*). Obviously there is much more to be said about this story, and the extent to which the 'key' analysis succeeds in highlighting stylistically important features. But the main point of this section of my paper has been to illustrate the potential of such analyses, using a chosen text and three alternative reference corpora of different generality.

4. Conclusion

In this article I have briefly explored a method of computer-aided stylistic analysis, involving the comparison of a focus text and one or more reference corpora. The technique is to employ the WMatrix software to identify and display items in order of keyness, or distinctiveness in the focal text, as contrasted with the reference corpus, measured in terms of the significance ratio of Log Likelihood. The main difficulty with this was the relative shortness of the 'The Mark', which gave undue prominence to some features occurring only a few times.

It is worthwhile, finally, noting some of the limitations as well as the future possibilities of this stylistic method. It is only too obvious, to begin with, that this type of analysis when applied to very large quantities of electronic text would be virtually impossible without the power of the modern computer. The great advantage of the techniques illustrated here is that they can be carried out automatically and at great speed. Wmatrix also shows great adaptability to the use of a wide range of corpora. The variety of corpora capable of being used is limited only by the user's ability to assemble the corpora and load them as 'personal folders' onto the Wmatrix website.

The corresponding disadvantage is that any activity involving human scrutiny of the data is immensely slow by comparison. Although POS tagging and semantic tagging are relatively accurate, there are still plenty of 'mistakes made by the computer' that ideally need to be manually checked. Further, although at present Wmatrix can operate with grammatical tags and semantic tags, there are many other levels of analysis that at present it cannot undertake – most importantly, parsing: the systematic syntactic analysis of a text in terms of phrases, clauses and so forth. There are also some more meaning-oriented stylistic analytic tasks (e.g. identifying metaphor or irony) that cannot (yet) be achieved by a computer.

The present situation, then, is that certain tasks can be undertaken fast but fallibly by computer, while other tasks can be undertaken more reliably but more slowly by human beings. Wmatrix already has the advantage that it can undertake a multi-level linguistic analysis of English corpora. Some of the items highlighted by the statistical analysis can clearly be seen to have thematic and literary significance, although without the help of WMatrix, they probably would not have been noticed.

One of the things suggested by this analysis is that there is no need to worry unduly about choosing an exactly appropriate reference corpus. None of the three reference corpora used in this experiment were ideal for the purpose, and yet the differences between the results of using the different reference corpora were rather minor.

Obviously this small experiment is far from exhaustive. I believe that present results, although lacking in detail, are promising, and that we can look forward to a future in which more revealing analyses of style can be achieved by computer at a more abstract level.

References

LEECH, Geoffrey (2008), *Language in literature: style and foregrounding*. Harlow: Pearson/Longman.

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

SCOTT, Mike (2004), *WordSmith Tools* version 4, Oxford: Oxford University Press.

WOOLF, Virginia (1917), 'The Mark on the Wall', in: Leonard Woolf and Virginia Woolf, *Two Stories*. London: Hogarth Press.