Stylistics: corpus approaches

Martin Wynne, Oxford University, 2005.

Introduction

Stylistics, which may be defined as the study of the language of literature, makes use of various tools of linguistic analysis. Corpus linguistics is opening up new vistas for the study of language, and there are interesting similarities in the approaches of stylistics and corpus linguistics. Stylistics is a field of empirical inquiry, in which the insights and techniques of linguistic theory are used to analyse literary texts. A typical way to do stylistics is to apply the systems of categorisation and analysis of linguistic science to poems and prose, using theories relating to, for example, phonetics, syntax and semantics. Theories and techniques of analysis from other areas such as sociolinguistics, pragmatics, cognitive linguistics and historical linguistics are also brought to bear on texts. These approaches are typified by Leech and Short (1981) and Short (1996).

The empirical approach to stylistics relies on linguistic evidence in the literary work. Corpus linguistics, also an empirical approach to linguistic description, relies on the evidence of language usage as collected and analysed in corpora. Burrows (2002) pointed out some underlying similarities of approach: "Traditional and computational forms of stylistics have more in common than is obvious at first sight. Both rely upon the close analysis of texts, and both benefit from opportunities for comparison." As linguists and stylisticians have become more aware of the possibilities offered by corpus resources and techniques, there have been increasing numbers of studies published which suggest that the coming together of these fields can be fruitful (e.g. Stubbs 2005). It is perhaps surprising, then, that apart from some important studies described here, there is little use of language corpora, or the techniques of corpus linguistics, in the study of literary style. But the use of language corpora is becoming more widespread and more mainstream, as various barriers to the their uptake are overcome. Among the important areas of current activity, at least two different approaches (corpus annotation and the analysis of collocation) can be discerned.

Corpus annotation

Corpus annotation involves investigating a particular linguistic feature by taking (or making) a corpus — a sample or complete collection of the texts to be studied in electronic form — and conducting a thorough and exhaustive analysis of the feature as it occurs in this corpus. The results of the analysis are normally

inserted into the electronic text as tags, or annotations. This activity of linguistic analysis and annotation of the text is similar to the procedures involved in word-class tagging, parsing and other forms of linguistic annotation, which are widespread activities. In this case we are interested more specifically in the annotation of literary texts, or in the annotation of discursive and stylistic categories (see also Leech *et al* 1997). Indeed, these stylistic annotations may make use of and build on syntactic tagging already inserted in the corpus.

There are typically three outcomes of this process. First, the exhaustive analysis of a whole text or corpus is a more empirically sound procedure for discovering linguistic phenomena, compared to choosing examples; annotation of the electronic text forces the analyst to test and refine the system of categorisation to account for all cases. Second, it is possible to extract statistics relating to frequency, distribution and co-occurrence of forms from the annotated text. Third, an annotated corpus is obtained, available for studies aiming to replicate or further develop the research, and usable for other areas of literary or linguistic research.

This approach is exemplified by the work done in the UK at Lancaster University on the forms of speech, thought and writing presentation in a corpus of texts. Leech and Short (1981) developed a system of classification for speech presentation in the novel. A project to test and refine this theory by attempting to apply it systematically to real data in a corpus was carried out over a number of years (Semino and Short 2004). A corpus of modern British English narrative texts was constructed, representing fiction, news reports and biographical writing. The corpus was then manually analysed such that each occurrence of any type of reporting, or presentation, of a language or thought event (e.g. direct speech, indirect speech, free indirect thought, etc.) was categorised and annotated in the corpus. This enabled the analysts to test the adequacy of the theoretical model against real data. It forced them to account for all relevant phenomena found in the texts, not just the interesting examples which they had chosen to retrieve. It also made possible qualitative and quantitative comparisons between the different text types. Among the findings of the project were the discoveries that it was necessary to adapt the model so that there were different scales for the presentation of speech, thought and writing, and that attempts to describe these phenomena together, as reported discourse presentation, risked missing the specificity of the presentation of the different modes. The distributions of the various forms were mapped across the text types, new categories were discovered, some categories were merged and numerous correlations between rhetorical function and stylistic choices were noted (Semino and Short 2004).

It should be noted that this project contrasted literary, non-literary narrative texts, and also contrasted high- and lowbrow texts within the genres. The assumptions underlying the category of "literature" and its sub-types are by no means fixed or uncontroversial, and the techniques and practices outlined above were also applied to non-literary texts and to texts whose "literariness" is under investigation. Using corpora, Carter (2004) has revealed linguistic creativity in everyday talk, and the work of the Lancaster stylisticians has continued with the analysis of spoken conversation (Semino and Short 2004).

Norms, deviations and collocations

A second approach which makes use of a corpus for stylistic research is to study literary effects in texts by using the evidence of language norms in a reference corpus. These effects can often be described as deviations from the norms of language use. The norms can be studied in a fairly straightforward way by looking in a large corpus. According to Stubbs (2005), “individual texts can be explained only against a background of what is normal and expected in general language use, and this is precisely the comparative information that quantitative corpus data can provide. An understanding of the background of the usual and everyday - what happens millions of times - is necessary in order to understand the unique.” So, for example, if a particular word or phrase (or a particular type of usage or meaning of a word or phrase) is thought to be an exclusively literary form, then it can be searched for by automatic or semi-automatic procedures in a corpus of non-literary texts in order to test this hypothesis.

A related area of increasing interest is the notion of 'semantic prosody'. Computational techniques can show patterns of co-occurrence of lexical items (collocations) and grammatical forms (colligations). Several corpus linguists have used evidence of these patterns to study creativity in language, both in fiction and in everyday usage (Sinclair, 1987; Carter, 2004; Hoey, 2005; Stubbs 2005), and the work of William Louw is of particular importance for stylistic studies.

Louw (1993), following on from the work of Firth and Sinclair, developed a new methodology for analysing literary effects through the study of collocations. The method is based on the idea that certain words, phrases and constructions become associated with certain types of meaning due to their regular co-occurrence with the words of a particular semantic category. To put it another way, the habitual collocates of a word give it a semantic colouring, which becomes part of the meaning of the word. For example Sinclair (1987) discovered that the subjects of the phrasal verb 'set in' are almost always unpleasant things (e.g “*rigor mortis* had set in”). This allows the possibility to evoke unpleasantness simply through

employing the phrasal verb, without using other evaluative words or phrases. Louw (1993) described how the word 'utterly' is used in this way in the Larkin poem *First Sight*, and developed a general theory of how the reader's feeling for semantic prosody can be exploited for ironic effect. Louw argued that an explanation or an analysis of the semantic prosodies associated with particular words is not generally accessible to our intuition. Such prosodies are essentially phenomena that can only be revealed computationally, and whose extent and development can only be properly traced by computational methods.

The application of the notions of collocation, colligation and semantic prosody are also being developed by the current work of Michael Hoey, whose theory of *lexical priming* adds a cognitive dimension and can be used to account for creativity in language (Hoey 2005). Speakers and hearers associate meanings with words not just because of their intrinsic meaning, but also because of the linguistic contexts in which they become habituated to speaking and hearing them. In this way words are *primed* for certain uses and meanings. For Hoey, creativity involves a selective overriding of the word's primings. The ways in which these primings are created by habitual usage can be found in corpora, and thus the source of the creativity can be studied in an empirical way. Interestingly, this work finds links not only between corpora and the study of style and creativity, but with cognitive aspects of language use as well.

Resources, tools and methods

If corpus linguistics and stylistics are so suited to each other in these ways, why is there not more work on the interface of these fields? Why do we still only talk about the potential for this area? There are several reasons why the potential for the use of corpora in stylistics has not been exploited to any large extent. For historical and institutional reasons practitioners of stylistics, with training in more traditional methods of humanities research, may not be skilled or equipped to use computers in their research. Furthermore, there is a lack of good quality, usable electronic texts and it is difficult to find and evaluate what is available. Though many texts can be found somewhere in electronic form, there is enormous variation in editorial principles, file formats, text encoding practices, documentation and quality control. This means that it is difficult to have confidence in the quality, consistency and integrity of many electronic texts. Users often need a high level of familiarity with text encoding, tagging schemes, text processing and text analysis software, along with an ability to deal with often complex generic computer hardware and software, in order to do the simplest things with texts on a computer. There is a lot of scope to develop textual resources and software to make research easier.

At a more philosophical level, there are trends of resistance to all of the more scientific, mathematical and empirical studies of literature, and the use of computers may seem to epitomise the non-literary and non-humanist approach to literature. This view is dramatised in David Lodge's novel *Small World* (1985), when a researcher reveals to a novelist the results of a statistical analysis of his style, and as a result the novelist is unable to write creatively again. Though stylisticians are not generally anti-scientific, there are some for whom computational procedures are a step too far. Although it can be argued that the use of computers for analysing electronic versions of texts, and for establishing evidence of linguistic norms in language use, is merely a means of verifying and refining empirical statements and findings, some see the danger of research becoming preoccupied with computational procedures, and the encoding and annotation of electronic texts, leading to a regrettable lack of attention to textuality and the meaning.

Willie van Peer warned of the dangers of looking at language out of context:

> "When stylistic features of a text have been transformed into numerical form, they acquire a status that actually prevents them from being perceived as language-for-communication as such. That is to say, in the very act of transforming textual qualities into counts, their essential process-like character is irretrievably lost. [...] Thus no level of (mathematical) sophistication is able to overcome the problem that the processes of meaning constitution have been eliminated before the analysis is undertaken. " (van Peer 1989: 302)

Similar criticisms have been voiced from within the field of corpus linguistics. Sinclair (2004), in particular, stressed the importance of not forgetting about the text and meaning.

Another important barrier to work with electronic texts is the fact that intellectual property rights, which aim to safeguard the rights of authors, present significant obstacles to academic research. In the real world (*pace* David Lodge) living writers are unlikely to be unexpectedly confronted with statistical analyses of their work, because making or copying an electronic version of their texts without permission is usually illegal.

Conclusion

As empirical work in linguistics increasingly makes use of language corpora, then stylistics and corpus linguistics are likely to continue to converge and overlap. Technical advances and improved resources are making the exploitation of electronic texts a more mainstream activity in stylistics. Large-scale projects to produce digital libraries of high quality texts are underway, and should overcome many of the current difficulties in finding reliable texts on the web. The establishment of important international standards

and guidelines for good practice in text encoding, such as XML, and the Text Encoding Initiative (TEI), are helping improve the quality and reliability of literary texts and corpora in electronic form. With standardisation of formats and procedures, we can hope for powerful, flexible and usable software tools for the analysis of literary texts and language corpora. Theoretical objections to the use of corpora in the study of the language of literature will doubtless remain, but it is to be expected that corpus linguistics will prove to be a useful addition to the stylistician's toolkit.

Bibliography

Burrows, J (2002). 'The Englishing of Juvenal: Computational Stylistics and Translated Texts.' *Style* 36:4, 677-679.

Carter, R. (2004). *Language and creativity: the art of common talk*. London: Routledge.

Hoey, M. (2005). *Lexical priming: a new theory of words and language.* London: Routledge.

Leech, G. N., McEnery A. M. & Wynne, M. (1997). 'Further levels of annotation'. In Garside, R. G., Leech, G. N. & McEnery, A. M. (eds.). *Corpus annotation*. Longman: Longman. 85-101.

Leech, G. N. & Short M. H. (1981). *Style in fiction: a linguistic introduction to English fictional prose*. London: Longman.

Lodge, D. (1985). *Small world: an academic romance*. Harmondsworth: Penguin.

Louw, W. (1993). 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies' in Baker, M., Francis, G. & Tognini-Bonelli, E. (eds.). *Text and technology*. Amsterdam: John Benjamins. 157-176. [Reprinted in Sampson, G. and McCarthy, D. (eds.) (2004). *Corpus linguistics: readings in a widening discipline*. London: Continuum. 229-241.]

van Peer, W. (1989). 'Quantitative studies of style: a critique and an outlook', *Computers and the Humanities* 23, 301-307.

Semino, E. & Short, M. H. (2004). *Corpus stylistics*. London: Longman.

Short, M. H. (1996). *Exploring the language of poems, plays, and prose*. London: Longman.

Sinclair, J. (1987). 'The nature of the evidence' in Sinclair, J. (ed.) *Looking up*. Glasgow: Collins, 150-159.

Sinclair, J. (2004). *Trust the text*. London: Routledge.

Stubbs, M. (2005). 'Conrad in the computer: examples of quantitative stylistic methods' in *Language and Literature*, 14:1.