Conrad in the computer: examples of quantitative stylistic methods

Michael Stubbs, *University of Trier, Germany*

Abstract

A stylistic analysis of Joseph Conrad's *Heart of Darkness* is used to illustrate the literary value of simple quantitative text and corpus data. Cultural and literary aspects of the book are briefly discussed. It is then shown that data on the frequencies and distributions of individual words and recurrent phraseology can not only provide a more detailed descriptive basis for widely accepted literary interpretations of the book, but also identify significant linguistic features which literary critics seem not to have noticed. The argument provides a response to scepticism of quantitative stylistics from both linguists and literary critics.

Keywords: *Conrad, Joseph; corpus stylistics; intertext; phraseology; quantitative methods*

This article applies quantitative methods of text and corpus analysis to a stylistic interpretation of Joseph Conrad's *Heart of Darkness*. I will try to bear in mind two criteria for computer-assisted methods which were set out very clearly by Kenny (1992): they must provide results which would be impossible to obtain without a computer, and they must be respected as an original scholarly contribution within literary studies. In the early 1990s, Kenny could find only a 'sadly small' number of studies which contain 'solid results obtained by techniques for which the computer is indispensable'.

1 Introduction

Stylistics has long led an uneasy half-life, never fully accepted, for many related reasons, by either linguists or literary critics. Linguists are often sceptical of stylistics because they are less interested in explaining particular individual texts than in developing general theories, and there is no convincing theory of text-types within which a theory of literary texts might be situated. However, 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. Literary scholars are often sceptical of linguistic description *per se*, and especially of apparently simplistic and

Language and Literature Copyright © 2005 SAGE Publications
(London, Thousand Oaks, CA and New Delhi), Vol 14(1): 5–24
DOI: 10.1177/0963947005048873 www.sagepublications.com

reductionist claims that statistics can define a literary style or contribute to a literary interpretation. Further, even if stylistic analyses of poems are accepted as complementing other methods of close reading, these methods seem unworkable for novels. However, individual hermeneutic methods also simplify, and quantification can make more explicit the evidence on which interpretations are based. Sometimes it is useful to reduce huge amounts of information to simple summaries ('All Jane Austen's novels are social satires about courtship and marriage'), but sometimes detailed statistics are required to reveal 'hitherto inaccessible regions of the language [which] defy the most accurate memory and the finest powers of discrimination' (Burrows, 1987: 3).

In a notorious attack on quantitative procedures, which has never been fully answered, Fish (1996) accuses stylistics of being 'circular' and 'arbitrary', of relying on selective attention to data, and of being caught in a logical dilemma. Either we select a few linguistic features, which we know how to describe, and ignore the rest; or we select features which we already know are important, describe them, and then claim they are important. Since a comprehensive description is impossible, and since there is no way to attach definitive meanings to specific formal features, stylisticians are apparently caught in a logical fork (which I will call the Fish Fork). Yet even if quantification only confirms what we already know, this is no bad thing. Indeed, in developing a new method, it is perhaps better not to find anything too new, but to confirm findings from many years of traditional study, since this gives confidence that the method can be relied on. I will return below to a very simple response to the Fish Fork: namely that it applies to any study of anything.

Scholars in other areas have also debated questions such as the duty of analysts to use all sources of information (including quantitative), the appropriate balance between a phobia of counting things and a naive confidence in technology, and the fact that quantification itself depends on prior subjective decisions and classifications (and that not everything is quantifiable). Several of these points are discussed not only in stylistics, but in work on quantitative methods in history (Jarausch et al., 1985) and in diachronic linguistics (McMahon and McMahon, 2003).

I will certainly not argue that a purely automatic stylistic analysis is possible. The linguist selects which features to study, the corpus linguist is restricted to features which the software can find, and these features still require a literary interpretation. However, since authors express their ideas through language, software can identify textual features which are of literary significance, including features which critics seem not to have noticed. In addition, as Sinclair (1975) argues, there is a serious gap in linguistic theories if they cannot explain the language of those texts which have the highest literary and cultural prestige.

2 Presentation conventions

Page references to *Heart of Darkness* are to the (identically paginated) Penguin Modern Classics (1973) and Popular Classics (1994) editions. All linguistic examples which are italicized or presented on separate lines are from this text. Frequencies of words and phrases are given in diamond brackets, e.g. <freq 10>. Lemmas (lexemes) are in upper-case, and word-forms are in lower-case: e.g. the lemma SEEM consists of the word-forms *seem*, *seems*, *seemed* and *seeming*. It is sometimes convenient to treat different parts of speech as members of the same lemma (e.g. *silence*, *silent*), and sometimes revealing to separate them (e.g. *knowledge* [noun] versus KNOW [verb]). This question is unresolved in lexical theory.

3 Background remarks

Joseph Conrad's *Heart of Darkness* is a very short novel, of less than 40,000 words. It was published in 1899 in magazine instalments, then in 1902 in book form. Carefully edited versions (e.g. Kimbrough, 1988; Hampson, 1995) identify autobiographical and other contemporary sources, and computer-readable versions are available (e.g. from Project Gutenberg).

A hundred years after its publication, it is still possibly 'the most commonly prescribed novel in [. . .] literature courses [. . .] in American universities' (Achebe, 1988). This raises the puzzle of why the text is still so popular, but might also imply that it is a poor choice for formal linguistic analysis. First, it seems unlikely that computer-assisted techniques could say anything new about a text which has been intensively studied for 100 years. Second, it is an overtly political text about late Victorian ideas of colonial power, which has been read from different political positions. It was famously attacked by Achebe (1988) for presenting stereotyped and racist views of Africans, and defended by others (e.g. Harris, 1988; Sarvan, 1988; Singh, 1988), who point out that the narrator Marlow describes at least the more unpleasant sides of colonial exploitation as a *sordid farce* and *senseless delusion* (p. 19). All this is a reminder that books can be interpreted in different ways, and that stylistic analysis will not provide a definitive reading.

Yet, despite extensive critical discussion, there is surprisingly little work on the book's linguistic style. Stampfl (1991) gives a psychological interpretation of frequent linguistic features. Erickson (2002) identifies syntactic features which appear to represent interference from French, Conrad's second language after his native Polish. Greaney (2002) writes about language as a theme in Conrad's books, but not about their language and style.

4 Narrative frames and main themes

The narrative of *Heart of Darkness* is embedded in different frames:

1. The book starts with an unnamed narrator on a boat on the Thames.
    2. Marlow becomes the narrator, and talks about the Thames in Roman times.
        3. Marlow tells of his visit to a European city.
            4. Marlow tells the story which takes up most of the book: he travels up a river in Africa in search of an ivory trader called Kurtz. He finds him, but Kurtz dies on the trip back down river.
        5. Marlow tells of his visit to Kurtz's fiancée back in the European city.
    6. [There is nothing corresponding to frame 2, but some vocabulary from frame 2 is repeated in frame 7.]
7. The book ends with a paragraph from the unnamed narrator back on the Thames.

Marlow's boat trip turns into his obsession with Kurtz, a trader who has been stealing ivory from the inhabitants of the region. He has apparently gone mad, is worshipped as a god by the native population, seems to have an African mistress, and may have been implicated in cannibalism.

Major places in the book are never named. We know that Conrad went to Brussels, where he arranged to travel to the Belgian Congo, and up the River Congo, but these place names never appear. Even the word *Africa* appears only once, when Marlow is discussing maps which he looked at as a child. Of course, if words do not appear in the text, then the computer cannot find them, and the reader has to infer them (from quite simple clues). The timescale is often also very vague: travelling up the river is *like travelling back to the earliest beginnings of the world* (p. 48); the story is told partly out of sequence, with flash-backs and flash-forwards; Marlow suddenly finds himself back in Brussels after a period of time which he remembers only mistily (p. 102), and with a few exceptions, people are not named, but identified by their functions (the *Lawyer*, Marlow's *aunt*, the *doctor*, the *manager*, Kurtz's *Intended*, etc). There are many other examples of this vagueness about places, times and people.

Given the extensive criticism of the book, we cannot approach it as naive readers (though the software can). There is considerable consensus among critics about major leitmotifs and themes, including the hypocrisy of the colonizers, and breakdown as a symbol of the unreliability of progress and civilization. Marlow's boat keeps breaking down, the colonial outposts are littered with *decaying machinery* (p. 22), Kurtz has a mental breakdown, and there are breakdowns in communication: people speak different languages, Marlow tells a lie about Kurtz to Kurtz's fiancée, and amongst the most frequent content words in the book is the lemma *SILENCE* <freq 37>. Other major themes are conveyed by repeated lexical contrasts, especially light and dark, restraint and frenzy, appearance and reality. There are frequent references to dreams (p. 48), nightmares (p. 100), trances (p. 56), phantoms and apparitions (pp. 85, 87, 105, 110) and visions (p. 105), and Marlow has trouble maintaining *contact with reality* (pp. 19, 54).

Critics point out that Conrad uses these contrasts to question whether 'heart of darkness' refers to 'darkest Africa', as the stereotype has it, or rather to the immorality of the white colonialists. In fashionable modern terminology, Conrad deconstructs the often taken-for-granted oppositions, white–black and good–bad. For readers around 1900, there would have been intertextual references to the books *Through the Dark Continent* and *In Darkest Africa*, published in 1878 and 1890 respectively, by Henry Morton Stanley (the Stanley who 'found' David Livingstone). The frequent contrasts of light and dark may also have recalled Genesis 1:1 to 1:5.

All of these points are clear in many literary critical discussions, so I now start to relate them to linguistic features of the text.

5 Vague impressions and unreliable knowledge

A further major theme of the book is Marlow's unreliable and distorted knowledge. Marlow himself never quite understands (and readers never quite find out) what *monstrous passions* (p. 95) and *vile desires* (p. 105) Kurtz has indulged. Conrad writes that an unnamed narrator says that Marlow says that an unnamed Russian says that Kurtz has talked to him: but we never discover what Kurtz has said. At the end of a series of story-tellers who quote story-tellers, nothing reliable remains. Kurtz dies uttering the words *The horror! The horror!*, but we never find out what that refers to: perhaps that Kurtz is now horrified by what he himself has done? (A useful project would be to apply to the book the detailed corpus-based categories of discourse presentation developed by Semino and Short, 2004.)

The unnamed narrator in the outside frame comments ironically that *we knew we were fated [. . .] to hear about one of Marlow's inconclusive experiences* (p. 10). Leavis (1962: 180) thought that Conrad simply didn't know what he wanted to say. (So did E. M. Forster [1936] cited by Haugh [1988].) However, Watt (1988) argues that lack of clarity is part of the point of this impressionist and early modernist story. He points out that mist or haze is a persistent image, and words from this lexical field are frequent in the book (in total almost 150, well over one per page on average):

- blurred 2, dark/ly/ness 52, dusk 7, fog 9, gloom/y 14, haze 2, mist/misty 7, murky 2, shadow/s/y 21, shade 8, shape/s/d 13, smoke 10[1], vapour 1

Marlow is frequently looking into a fog, uncertain of what he is seeing. Things are constantly *in a muddle* and *chaos* (p. 26). Unexplained things happen: a man hangs himself for no apparent reason (p. 21). As he approaches Kurtz, Marlow is confused by signs and symbols that he cannot decode: a faded message on a board which is difficult to *decipher* (he cannot read its *illegible signature* at all, p. 53); a book with strange writing in the margins which he thinks is *cipher* (it turns out to be Russian, pp. 54, 78); round carved balls on posts (which turn out to be *symbolic*: when he looks at them more carefully through a telescope, he sees that they are human skulls, pp. 75, 82); and a figure dressed as a harlequin (p. 75), who is *improbable, inexplicable, and altogether bewildering* (p. 78).

Literary critics tend to identify a few content words, such as *fog* and *mist*, *vague* <5> and *indistinct* <4>:

- I saw vague forms of men
- Marlow ceased, and sat apart, indistinct and silent

However, they tend to ignore the many grammatical words denoting vagueness and uncertainty. The word *something* occurs over 50 times, in expressions such as:

- I don't know – something not quite right
- reminded me of something I had seen – something funny

There are over 200 occurrences of *something*, *somebody*, *sometimes*, *somewhere*, *somehow* and *some*, plus around 100 occurrences of *like* (as preposition), plus over 25 occurrences of *kind of* and *sort of*, all often collocated with other expressions of vagueness:

- the outlines of some sort of building
- seemed somehow to throw a kind of light
- I thought I could see a kind of motion
- indistinct, like a vapour exhaled by the earth . . . misty and silent

If we add all this to occurrences of *seemed* <ca 50>, words expressing vagueness are very frequent <ca 385>: well over three per page. Here are simple comparative frequency statistics for vague words, normalized to occurrences per 1000 running words.

(1) HEART = *Heart of Darkness*.
(2) FICTION = a corpus of fictional texts of over 710,000 words.
The reference corpus here was the 'imaginative prose' categories K, L, M and N in Brown, LOB, Frown and FLOB.
(3) WRITTEN = the one-million-word written component of the BNC sampler.[2]

	(1) HEART	(2) FICTION	(3) WRITTEN
some	2.6	1.5	1.5
something	1.3	1.0	0.4
somebody	0.2	0.1	0.05
sometimes	0.6	0.2	0.2
somewhere	0.2	0.2	0.03
somehow	0.2	0.1	0.04

Frequencies are consistently higher in HEART than in FICTION, and higher in FICTION than in WRITTEN.

6 Some (very) simple frequency data

Textual frequency is not the same as salience, and does not necessarily correspond to what readers notice and remember in a text. This is one clear limitation on studies which look only at the words on the page, and I will show below that word-frequency lists have other severe limitations. Neverthless, there must be some relation, even if indirect, between frequent vocabulary and content, and frequency lists are one essential starting point for a systematic textual analysis.

A list of the most frequent lexical words (i.e. excluding very high frequency grammatical words) in *Heart of Darkness* does not initially look very promising:

- said 131, like 122, man 111, Kurtz 100, see 92, know 87, time 77, seemed 69, made 65, river 65, came 63, little 62, looked 56, men 51, Mr 51, long 50

Using 'keywords' software (Scott, 1997), we can also check which words are both frequent in *Heart of Darkness* and also significantly more frequent than in a reference corpus (again the 'imaginative prose' corpus defined above).[3] These are all the content words among the top 50 keywords, which both occur 20 times or more in the novel and are also significantly more frequent than in the reference corpus (listed in descending frequency in the novel):

- Kurtz 100, seemed 69, river 65, station 48, great 46, manager 42, earth 39, ivory 31, pilgrims 31, darkness 25, bank 25, forest 23, wilderness 22, Kurtz's 21, cried 20

Frequent nouns may indicate superficial topics in a text (*Kurtz*, *river*), but not its underlying themes: this is not a book 'about' a river. Verbs are often a better candidate for stylistically relevant words. So, we can also use software to lemmatize the text and list the top 10 verb lemmas:

- SAY, SEE, LOOK, KNOW, COME, MAKE, SEEM, HEAR, TAKE, THINK

Now, the most frequent word-form of all (*said*) and the most frequent verb lemma (SAY) are of course very frequent in fiction in general, and the mental verbs (*see*, *know*, *looked*, etc.) are also usually frequent in fictional texts (for quantitative data, see Stubbs and Barth, 2003). However, other words are of more interest: many occurrences of *like* <ca 100> and of *looked* <ca 25> are in vague expressions such as 'x was like y' and 'x looked like y' or 'it looked as though'. This implies that we must look not just at individual words, but at their recurrent phraseology: see below.

Although such lists might seem to offer only the crudest kind of content summary, note for the present that SEEM is among the top words in all three lists, and that several words in the lists concern uncertainty, perception and knowledge. This statement clearly involves a subjective interpretation of potential literary significance, but the statement is based on objective textual features.

7 Word distribution and text structure

The first reason why counting individual words is certainly not sufficient is that the interesting content words are not (by definition) evenly distributed across the text, but are clustered at different places. For example, some words (e.g. *Buddha*) occur only in the opening and closing narrative frames (pp. 10, 111) and are thus used to mark text structure. The book starts on the Thames, describes a journey up the River Congo (unnamed), and ends back on the Thames. At the beginning and the end we have:

- the Thames . . . a waterway leading to the uttermost ends of the earth (pp. 5–6) . . . Marlow sat cross-legged . . . he had the pose of a Buddha . . . we felt meditative (pp. 6, 10).
- Marlow . . . sat apart . . . in the pose of a meditating Buddha (p. 111) . . . the tranquil waterway leading to the uttermost ends of the earth (p. 111, last sentence).

At the beginning, Marlow visits a city like a *whited sepulchre* (p. 14). He enters offices, past *high houses* in a *narrow and deserted street*, through *doors standing ponderously ajar*. In the offices *a door open[s]* (p. 14). At the end, back in the *sepulchral city* (p. 102), he visits Kurtz's Intended through a *ponderous door, between tall houses* in a street as quiet as *a cemetery*. In the Intended's house is a piano like a *sarcophagus*. A *high door open[s]* and the Intended comes in (p. 105). This verbal trick is also used to express Marlow's similarities to Kurtz: both are described as a *voice* (pp. 39, 67, 69, 86) and as an *idol* (pp. 6, 84).

Such distributional facts, which start to say something about the structure of the whole text, can be tracked with a simple program. For example, the words *heart*, *dark* and *darkness* occur throughout the book, but increase in frequency at the very end when the story almost becomes *too dark – too dark altogether* (p. 111). Similarly, the lemmas DREAM <15> and NIGHTMARE <6> are very differently distributed. DREAM occurs twice at the very beginning, then several times in a cluster, when Marlow is *trying to tell* his dream (p. 39), then fairly regularly throughout the rest of the story. NIGHTMARE occurs once at the beginning, where there are *hints for nightmares* (p. 21), and then in a cluster towards the end (ca p. 95), all in collocations with *Kurtz*. In terms of word distribution, Marlow's dream turns into a nightmare.

The verb lemma KNOW is frequent <122>, and fairly evenly distributed throughout the text. Many instances are negative, either grammatically (*I don't know*, *he did not know*) or by implication (*he wanted to know*, *if only he had known*). This is a novel about the fallibility and distortions of human knowledge. Right at the end, there is a cluster of positive examples: Kurtz's Intended repeats over and over that

- I alone know how to mourn for him . . . I am proud to know I understood him better than any one . . . You have heard him! You know! . . . You know what vast plans he had.

The irony is of course that she knows nothing of what Kurtz has done. (On how patterns of lexico-grammar can express the 'fallibility of human knowledge', see also Hardy and Durian, 2000.) These examples show how the distribution of individual words can be studied, but require (as Fish [1996] would doubtless point out) that the analyst knows beforehand which words might be of interest.[4]

8 Text and inter-text

The second reason why studying only individual words in the text is inadequate is that any text makes references to other texts. Allusions to the Aeneid, to the figures of the Fates in Greek mythology, to Dante's depiction of hell, and to the Faust legend are discussed in detail by Lothe (2000, 2001). *Heart of Darkness* is based explicitly on the metaphor of life as a journey. It contains elements of a parable, an adventure story, a mystery story, and a fairy tale: characters are *bewitched* (pp. 33, 39), held *captive by a spell* (p. 50); and Marlow's approach to Kurtz is *beset by [...] dangers as though he had been an enchanted princess sleeping in a fabulous castle* (p. 61).

Fools and jesters engage in absurd or incomprehensible activities and provide elements of black comedy and carnival: the clerk who claims he is *not such a fool as [he] looks* (p. 16); the *harmless fool* (p. 16) of a doctor who measures Marlow's skull; the man like *a hairdresser's dummy* with *a green parasol* (pp. 25–6); the brickmaker who makes no bricks (p. 34); the *papier maché Mephistopheles* (p. 37); the *little fat man* with *red whiskers* and *pink pyjamas* (p. 57); and, most explicitly, the Russian (*I am a simple man* [pp. 84, 90]) dressed as a *harlequin* (pp. 75, 78).

The frequent references to fools <*fool*, *foolish* 18> and to madness and insanity <*mad*, *madness*, *insanity* 11> can also be taken as allusions to the Ship of Fools, the medieval satire on vices and follies. (Sebastian Brandt's *Narrenschiff* of 1494 was adapted by Alexander Barclay in 1509 as *The Shyp of Folys of the Worlde*, and by W. H. Ireland in 1807 as *Modern Ship of Fools*.) The French steamer which takes Marlow to Africa shells the bush, with *a touch of insanity* (p. 20), in the belief that it is attacking enemies; and Marlow's steamboat has a *mad helmsman* (pp. 63, 64, 65, 73). As critics have pointed out (e.g. Dorall, 1988; LaBrasca, 1988), the themes of carnival, insanity and the ship of fools are perhaps even more evident in *Apocalypse Now*, Coppola's film adaptation of Conrad's book.

Quotes and near-quotes from specific texts can quickly be identified by computer-assisted searches. The phrase *whited sepulchre* (p. 14), with two later references to the *sepulchral city*, is from Matthew 23: 27. Other Biblical references were probably more obvious to readers around 1900. In the opening and closing paragraphs (pp. 5–6 and 111) of the book, we have the phrase

- waterway leading to the uttermost ends of the earth

The word *uttermost* is not frequent in general English (or in Conrad's other writings), but it is frequent <29> in the King James translation of the Bible, sometimes in the phrase *uttermost part(s) of the earth* <5>. The phrase *the ends of the earth* is also frequent <30> in the Bible. I will come back to such abstract place terms later.

There are lexical allusions to Dickens's *Tale of Two Cities* (the knitters of black wool recall the women knitting at the foot of the guillotine), and to Jules Verne's *Voyage au Centre de la Terre* (Marlow sets off as if on a journey to *the centre of the earth* [p. 18]; he is a *wanderer on a prehistoric earth*, as though the earth was *an unknown planet* [p. 51]).[5] Cannibalism and atavism (the fear that 'civilized' humans could revert to a more primitive type) were something of an obsession in Victorian Britain, and crop up in various pseudo-anthropological books of the time and in novels such as R. L. Stevenson's *The Strange Case of Dr Jekyll and Mr Hyde* (1886), H. G. Wells's *The Time Machine* (1895), Bram Stoker's *Dracula* (1897) and Conan Doyle's *The Hound of the Baskervilles* (1902), all published within a few years of *Heart of Darkness* (Griffith, 1995; Breuer, 1999). Intertextuality, by definition, implies that a text can be read at different levels. The women knitting black wool can be taken as merely a realist detail, but also as a reference to the Greek Fates and/or a reference to Dickens's women under the guillotine. The forest can be taken as merely an African forest and/or a reference to all the forests in which characters in folk-tales get lost.

9 Collocations: words and words

The third reason why individual words can only be a starting point is that collocations create connotations. For example, *grass* <freq 18> is usually associated with death, decay and desolation: it sprouts through the stones in the *city of the dead* (= Brussels, p. 14), and through the bones of a dead man (p. 13); old machinery is abandoned in it (p. 22). Here are selected examples (emphasis added):

ast to meet my predecessor, the	grass growing through his ribs was
enetian blinds, a dead silence,	grass sprouting between the stones,
upon a boiler wallowing in the	grass, then found a path leading up
er the empty land, through long	grass, through burnt grass, through
ically childish in the ruins of	grass walls. Day after day, with th
layer of silver – over the rank	grass, over the mud, upon the wall
as though he had been a wisp of	grass, and I saw the body roll over
mit was half buried in the high	grass; the large holes in the peake
all fours – I've got him.' The	grass was wet with dew. I strode ra

The words GLITTER <14>, GLEAM <8>, GLISTEN <3> and GLINT <2> connote things which are ominous and dangerous: GLITTER collocates with *dark*, *sombre*, *gloom*, and the *infernal stream*; GLEAM collocates with *blood* and *fire*; people's eyes *glitter*, *glisten* and *gleam*; arrows *glint* when they are being shot at Marlow (p. 65). Some such associations are signalled explicitly in the text. In Brussels, Marlow looks at the coloured patches on a map of colonial countries (p. 14). Later, he meets the Russian harlequin figure whose clothes are covered with coloured patches (p. 75), and Marlow says:

- His aspect reminded me of something I had seen – something funny I had seen somewhere.

The connotations of individual words can be inferred from their recurrent collocates, but only if we know in advance which words to look at (the Fish critique again). However, software can identify clusters of words which co-collocate across the text, by recording the collocates of each word in the text within a given span, and then summing the collocates for each node-word. (On this technique, see Phillips, 1985.) For example, 10 high frequency words (*the, of, and, to, a, in, that, is, was, it*) were deleted from the text, and collocates of all other words were recorded in a span of 10. One example of a collocational cluster which the software then identified is: *gloom*, *brooding*, *still/stillness*, *reaches* (in the sense of *the upper reaches of the Thames*), plus more marginal collocates: *sombre*, *death*, *black*, *silence*, *mysterious*.

10 Phraseology: words and grammar

A fourth limitation on looking at individual words is that words occur in recurrent lexico-grammatical patterns. Critics have complained that the book is repetitive: 'a bombardment of emotive words, and other forms of verbal trickery' (Achebe, 1988). However, *Heart of Darkness* is not very repetitive in the sense of using the same words from a small vocabulary over and over again. We can check how many different words Conrad uses as a proportion of the total running words (that is, the type–token ratio of the text). Over the first 2,000 words, this ratio rises a little more slowly than *Middlemarch* by George Eliot, a little faster than *Oliver Twist* by Charles Dickens, and considerably faster than *Death in the Afternoon* by Ernest Hemingway, who is almost stereotypically a writer who (deliberately) uses a small vocabulary. These comparative figures are provided by Youmans (1990). So *Heart of Darkness* is well within the norms of English fiction, and this does not explain complaints about repetitiveness.

The impression of repetitiveness arises rather from Conrad's use of particular grammatical patterns, including long strings of adjectives and nouns:

- the air was warm, thick, heavy, sluggish (p. 48)
- their glance was guileless, profound, confident, and trustful (p. 106)

- joy, fear, sorrow, devotion, valour, rage – who can tell? (p. 52)
- was it superstition, disgust, patience, fear [. . .]? (p. 60)

He also repeatedly uses nominal groups consisting of an abstract noun (usually two abstract nouns) plus an adjective with a negative prefix:

the aspect	of an unknown	planet
the darkness	of an impenetrable	night
the extremity	of an impotent	despair
the heart	of an impenetrable	darkness
the sea	of inexorable	time
the shape	of an unrestful and noisy	dream
the stillness	of an implacable	force
the test	of an inexorable physical	necessity

When Leavis (1962) famously complains about Conrad's repetitive style, he gives the individual words *inscrutable*, *inconceivable* and *unspeakable*, but seems to miss the grammatical generalization that Conrad uses a large number of words with negative prefixes <ca 200>: two per page on average. The following are the most frequent:

- impossible/ity 12, uneasy/iness 8, unexpected/ness 7, impenetrable 6, inconceivable/ly 6, incredible 5, indistinct/ly 5, intolerable/ly 5, unknown 5, incomprehensible 4, inscrutable 4, unearthly 4, unsound 4

In addition to these 200 or so words (mainly adjectives), a further 50 end in *-less* (e.g. *colourless*, *heartless*), and there are a further 500 occurrences of *no*, *not*, *never*, *nothing*, *nobody* and *nowhere*, plus a further 50 occurrences of *without*. The total frequency of these negatives is over 800: around one in every 50 words of running text.

11 On interpreting patterns

Now, the difficulty, to which Fish (1996) correctly points, is to say what this pattern means. A plausible interpretation is based on presuppositions. A negative statement usually implies that a positive was expected, and many examples emphasize how alien Africa is when contrasted with things back home. If Marlow says the African coast is *featureless* (p. 19) and *formless* (p. 20), this is because we do not expect coasts to be like this. If he says *there were no villages* (p. 58) along the river, then this is because we expect villages along rivers. If the river flows by *without a murmur* (p. 38), then this is because we expect a river to make some sound. If he says that *nothing happened* (pp. 20, 111), then this is because we expect something to happen. This theme is sometimes explicit: the earth is *unearthly*, as opposed to what we are *accustomed to look upon* (p. 51); and Marlow fails to recognize the skulls on the stakes because he *had expected to see a knob of wood* (p. 82). (EXPECT <17>)

Watt argues that 'there are no negatives in nature, but only in the human consciousness' (1960: 259). Tabata (1995: 102), quoting Watt, argues that negatives signal subjectivity, and shows that Dickens's first-person narratives have many more negatives than his third-person narratives. Werth (1995) argues that the negatives in the opening of E. M. Forster's *Passage to India* signal unexpected aspects of the scene. Hidalgo-Downing, quoting Werth, discusses the pragmatic function of negatives in implying more than is literally said (2000: 217): they deny expectations and challenge background propositions (2000: 223), they are a way of questioning reality and are therefore an alienation device (2000: 219, 222). Many of her comments apply to *Heart of Darkness*, where the frequent negatives represent a world which is strange, foreign, alien, contrary to cultural expectations, and *impenetrable to human thought* (p. 79). They construct a contrast between the supposedly civilized world which Marlow has left and the supposedly primitive world which he encounters. In 'An Outpost of Progress' (1898), a short story which is a precursor to *Heart of Darkness*, Conrad writes of the *vast and dark country* where the story takes place as *the negation of the habitual* and *the affirmation of the unusual*.

An odd footnote to this is provided by Ford Madox Ford (quoted in Kimbrough, 1988: 213), who tells of his discussions with Conrad about the negative sentence in the last paragraph of *Heart of Darkness*: *Nobody moved for a time* (p. 111). He states 'the first principle of the technique' of Conrad and himself at that time: '. . . you should never state a negative. If nobody moves, you do not have to make the statement; just as, if somebody is silent, you just do not record any speech of his, and leave it at that.' On the face of it, it would seem that Ford Madox Ford had completely misunderstood Conrad's technique.

Here are comparative frequency statistics for negatives (estimate per 1000 running words).

	(1) HEART	(2) FICTION	(3) WRITTEN
not	6.3	4.7	4.2
n't	4.2	6.3	1.7
no	4.1	3.2	1.9
never	1.0	1.2	0.5
nowhere	0.2	0.04	0.02
nothing	1.2	0.7	0.3
without	1.3	0.6	0.4

There are almost always more occurrences in HEART than in FICTION, and more in FICTION than in WRITTEN. Taken together, the frequency of *not* and *n't* is almost identical in HEART and FICTION, but still higher than in WRITTEN.

12 More on phraseology: recurrent word sequences

A fifth limitation of looking at individual words is that words occur in recurrent two-, three-, four- and five-word lexico-grammatical patterns which pervade the text. The top two-word sequence is of course a pair of grammatical words, *of the* <241>, which might seem of little interest, though a concordance shows that around half of the occurrences are followed by a place term: *of the forest(s)* <9>; *of the land* <9>; *of the river* <7>; *of the earth* <6>; *of the wilderness* <5>; *of the world* <5>; *of the stream* <4>; etc. This is three to four times higher than in a large corpus of general English.

The top two-word sequence which contains a content word is *seemed to* <46>: it occurs every couple of pages on average. Now *seemed* has risen right to the top of a frequency list.

	(1) HEART	(2) FICTION	(3) WRITTEN
seemed to	1.2	0.4	0.09

The top four-word sequences relate to the central themes of geographical and psychological space, appearance and reality, and Marlow's uncertainty about everything. The complete list of four-word sequences which occur more than five times each is

it seemed to me	7
as far as I	6
as though I had	6
with an air of	6
the depths of the	6

The first four concern uncertainty and the fifth is a place expression. We can also identify more abstract phrasal frames. These sequences each occur individually more than once, and differ in only one word from other sequences:

the bottom	of the	
the depths	of the	
the edge	of the	
the face	of the	
the middle	of the	
the midst	of the	
the recesses	of the	
the rest	of the	<total freq 26>

as though I	had	
as though he	had	
as though it	had	
as though they	had	<total freq 15>

I don't know how	
I don't know I	
I don't know what	
I don't know why	<total freq 10>

13 Appearance and reality, centre and periphery

These phrasal patterns both contribute to the feeling that the text is very repetitive, and also convey major themes in the text, such as Marlow's uncertainty about appearance and reality, and places, geographical and psychological. These themes are conveyed not only by words from specific lexical fields (e.g. 'fog' and 'dreams'), plus individual vague words, but also by these recurrent phrases. The observation that *Heart of Darkness* contains many negatives and many occurrences of *as if* and *as though* has been made by literary critics (Senn, 1980; Stampfl, 1991). I am claiming not that quantitative corpus methods produce entirely new insights into the text, but that they describe more accurately the range of lexico-grammatical patterns which Conrad uses.

So, bearing in mind the title of the book and its very last phrase

- the tranquil waterway [= the Thames]... seemed to lead into the heart of an immense darkness

I will look finally at the repeated phrases which have been picked out by a mechanical method, but which still require interpretation, such as

the depths	of the
the edge	of the
the middle	of the

The book contains a large number of phrases with the structure (PREP) the NOUN of DET NOUN, where the first noun is a place term. The following are only some examples.

	the	beginning	of an	interminable waterway
	the gloomy	border	of the	forest
beyond	the	bounds	of	permitted aspirations
	the very	brink	of the	stream
to	the	centre	of a	continent
into	the gloomy	circle	of some	inferno
in	the uttermost	depths	of	despondency
into	the	depths	of	darkness
in	the	depths	of the	land
on	the	depths	of the	sea
towards	the	depths	of the	wilderness
to	the	edge	of the	forest
	the very	end	of the	world
to	the uttermost	ends	of the	earth
into	the	heart	of	darkness
down	the	middle	of the	river
in	the	midst	of the	incomprehensible
from	the	recesses	of the	land

	the	skirts	of the	unknown
on	the	threshold	of	great things

There are over 170 occurrences of the construction *the * (*) of a/an/the*, such as

- the face of the forest; the glitter of the reach; the midst of a shocking hullabaloo; the long shadows of the forest.

14 Text and corpus

Corpus techniques open the possibility of systematic comparisons between individual texts and reference corpora, but comparative statements must be carefully checked. Although these place expressions have a particular significance in *Heart of Darkness*, they are not unusually frequent in comparison with general usage. On the contrary, the most frequent five-word phrases in the 100-million word British National Corpus (with frequencies per million running words) include:

- at the end of the 45; by the end of the 18; in the middle of the 16; at the top of the 11; at the beginning of the 9; at the bottom of the 7; on the edge of the 7; towards the end of 5; in the centre of the 6.

General corpus data show that people frequently talk about space (geographical, social and psychological), and especially about the centre and periphery of things. However, Conrad's place expressions differ in three ways from their frequent use in general English. First, they are all abstract and extremely vague, and (apart from the Thames) they could not be found on any map. They refer to an unnamed *forest* or *river*, to a mythical *inferno*, to an abstract *wilderness*, or to an even more abstract *darkness*. Second, they acquire evaluative connotations in context, in uses such as

- I had peeped over the edge (p. 101)
- he had stepped over the edge (p. 101)

General corpora show that *over the edge* has clear connotations of danger. Frequent collocates are *PUSH*, *hanging*, *GO*, *dangling*, *falling*, *dropped*, *tipping*, *dangerously*, *slipping*.[6]

Third, these recurrent phrases are read differently because of the title of the book, the repeated references to characters who are shades and phantoms, on the threshold between this world and the next, and the last sentence which contains two such phrases:

- to the uttermost ends of the earth
- into the heart of an immense darkness.

15 Conclusions

An unsolved problem for stylistics is how close attention to a text can be reconciled with an understanding of its cultural and historical background. My discussion assumes that a literary text is not a self-contained autonomous object. First, all texts consist of fragments of other texts: they make references to text-types (e.g. black comedy), to other stories (e.g. the Faust legend) and to individual texts (e.g. the Bible). Second, there are no clear boundaries between a literary text and general language use. Since a text is a selection from the potential of the language, we require hermeneutic methods which identify observable evidence of meaning in the form of inter-textual relations between texts and corpora. Comparative corpus methods now allow us to study how far texts consist of recurrent phrasal patterns which are widespread in the language as a whole. In some ways, the language of *Heart of Darkness* deviates from the norm of everyday language use, but many recurrent phrases in the book are significant because they exploit the routine phraseology of the language.

An overall discourse schema in the book is Europe and Africa, the River Thames and the River Congo, light and dark, all with the much commented ambiguity between these poles:

- 'And this [= the Thames] also', said Marlow suddenly, 'has been one of the dark places of the earth'. (p. 7)

These themes are conveyed partly by recurrent phrasal schemas which position characters and readers – depending on who they are – in the centre, on the brink, over the edge, or beyond the pale. These phrases can be used concretely and literally, but they constantly evolve – in texts and over time – into abstract and metaphorical uses, and this seems to be a linguistic universal. Does this help to explain why the book is so popular? It not only fits into widely popular text-types (e.g. adventure story) and contains repeated images from folk tales (e.g. dark forests and devils dancing by firelight) but also uses a highly frequent phraseology which reflects social obsessions and stereotypes of civilized and primitive, home and abroad, us and them, centre and periphery.

My argument might now seem to be impaled on the Fish Fork. If I had discovered that this phraseology was more frequent in *Heart of Darkness* than in general English, then I would have argued that it is interesting for *this* reason, but since I have discovered that it is a pervasive feature of general English, then I have argued that it is interesting for *that* reason. However, as I have noted above, I think there is a very simple response to the Fish Fork: it applies to any study of anything. Pure induction will never get you from empirical observations to interesting generalizations. You have to know where to look for interesting things. As Grice (1989: 173) puts it: 'you cannot ask [...] what something is unless (in a sense) you already know what it is'. However, this is true only 'in a sense', since the aim is to say systematically and explicitly what something is: and that is where empirical, observational analysis can contribute. It is not possible (or desirable) to avoid subjectivity, but observational data can provide more systematic evidence for unavoidable subjective interpretation.

The computer does not provide a single method of text analysis, but offers a range of exploratory techniques for investigating features of texts and corpora. The findings of corpus stylistics (comparative frequencies, distributions and the like) sometimes document more systematically what literary critics already know (and therefore add to methods of close reading), but they can also reveal otherwise invisible features of long texts. The phraseology which I have described is a formal, observable, objective feature of the book. It is only one feature, and it is open to different interpretations, but it was not created by my analysis.

Acknowledgements

This article is a revised version of the Sinclair Lecture, given at the University of Birmingham in May 2003, and published by the University of Birmingham in 2004 (Stubbs, 2004). The article was a long time in the writing, and I am very grateful to several seminar audiences and to the following friends and colleagues who all made helpful critical comments on earlier drafts: Horst Breuer, Andrea Gerbig, Gabi Keck, Anja Müller-Wood, John Sinclair, Bettina Starcke, Katrin Ungeheuer, Jean Jacques Weber and two anonymous referees. Isabel Barth wrote the software for identifying recurrent phrases, for tracing the distribution of words in a text, and for identifying collocational clusters. Bill Fletcher provided me with a beta version of his N-gram software for identifying repeated phrase frames (Fletcher, 2002).

Notes

[1] This is s*moke* in the relevant sense, excluding s*moking my pipe*, etc.!

[2] Brown, LOB, Frown and FLOB are parallel reference corpora of one million words each of written English. The Brown corpus (prepared at Brown University under the direction of W. Nelson Francis and Henry Kučera) comprises written American English, published in 1961. LOB (prepared at Lancaster, Oslo and Bergen Universities, under the direction of Geoffrey Leech and Stig Johansson) comprises written British English, published in 1961. Frown and FLOB (prepared at Freiburg University under the direction of Christian Mair) are comparable corpora, of written American and British English, published in 1991. The BNC sampler consists of one million words each of spoken and written British English, extracted from the 100-million-word British National Corpus.

[3] The 'keywords' tool is one part of the WordSmith Tools software (Scott, 1997). Given two word-frequency lists, for a given text and for a reference corpus, the software compares the two lists and finds words which are significantly more (or less) frequent in the text than in the corpus. Two statistics, chi-square and log-likelihood, can be used for the comparison. The log-likelihood statistic was used in this article. The reference corpus can contain any sample of texts which are judged to provide a useful comparison, but this sample will generally (a) be much larger than the text under analysis, and (b) either contain texts of the same genre or be a large mixed general corpus.

[4] Youmans (1991) defines a technique of textual analysis which tracks the changing distribution of all words in a text. He shows how an increase in the type–token ratio reveals the introduction of clusters of new words into a text, and how this may provide evidence of the structure of a narrative. Stubbs (2001: 123–44) replicates Youmans's analysis of a short story by James Joyce and adds some analytic modifications. Youmans's software is available at http://web.missouri.edu/~youmansc/vmp/index.shtml (last accessed 7 June 2004).

[5] Dickens's novel was published in 1859. It contains phrases such as: . . . *the Guillotine. In front*

of it . . . a number of women, busily knitting. . . . the women sat knitting, knitting. Darkness encompassed them. Another darkness was closing in as surely . . . the women sat knitting, knitting . . . knitting, knitting, counting dropping heads. Verne's novel was published in 1864 in French (and a few years later in English translation). It contains phrases such as: *to the very centre of the earth . . . a vision of the prehistoric world . . . heavy gloom of deep, thick, unfathomable darkness*.

6 One of the main findings of corpus analysis is that many words have evaluative connotations which signal speaker attitude. Louw (1993) is an influential article which shows how concordance data on frequent collocation provide observable evidence of pragmatic meanings. Many subsequent studies have confirmed the method and the findings with a wide range of examples. Stubbs (2001) discusses the concept and gives references to work by Bublitz, Channell, Church, Clear, Hanks, Hunston, Krishnamurthy, Moon, Partington, Sinclair, Tognini-Bonelli, and others.

References

Achebe, C. (1988) 'An Image of Africa: Racism in Conrad's *Heart of Darkness*', in R. Kimbrough (ed.) *Joseph Conrad: Heart of Darkness*, pp. 251–62. New York: Norton. (Originally published 1975.)

Breuer, H. (1999) 'Atavismus bei Joseph Conrad, Bram Stoker und Eugene O'Neill', *Anglia* 117(3): 368–94.

Burrows, J. F. (1987) *Computation into Criticism*. Oxford: Clarendon.

Dorall, E. N. (1988) 'Conrad and Coppola: Different Centres of Darkness', in R. Kimbrough (ed.) *Joseph Conrad: Heart of Darkness*, critical edn, 3rd edn, pp. 301–11. New York: Norton. (Originally published 1980.)

Erickson, J. L. (2002) 'Strict Adjacency and the Prose of Joseph Conrad', in S. Scholz, M. Klages, E. Hantson and U. Römer (eds) *Context and Cognition*, pp. 75–82. Munich: Langenscheid-Longman.

Fish, S. E. (1996) 'What is Stylistics and Why are They Saying Such Terrible Things About It?', in J. J. Weber (ed.) *The Stylistics Reader*, pp. 94–116. London: Arnold. (Originally published 1973.)

Fletcher, W. H. (2002) N-Gram Software. Available at http://kwicfinder.com/kfNgram/.

Greaney, M. (2002) *Conrad, Language, and Narrative*. Cambridge: Cambridge University Press.

Grice, H. P. (1989) 'Postwar Oxford Philosophy', in H. P. Grice *Studies in the Way of Words*, pp. 171–80. Cambridge, MA: Harvard University Press. (Originally published 1958.)

Griffith, J. W. (1995) *Joseph Conrad and the Anthropological Dilemma*. Oxford: Clarendon.

Hampson, R. (ed.) (1995) *Joseph Conrad: Heart of Darkness with Congo Diary*. Harmondsworth: Penguin.

Hardy, D. and Durian, D. (2000) 'The Stylistics of Syntactic Complements: Grammar and Seeing in Flannery O'Connor's Fiction', *Style* 34: 92–116.

Harris, W. (1988) 'The Frontier on which *Heart of Darkness* Stands', in R. Kimbrough (ed.) *Joseph Conrad: Heart of Darkness*, critical edn, 3rd edn, pp. 262–68. New York: Norton. (Originally published 1981.)

Haugh, R. H. (1988) '*Heart of Darkness*: problem for critics', in R. Kimbrough (ed.) *Joseph Conrad: Heart of Darkness*, critical edn, 3rd edn, pp. 239–42. New York: Norton. (Originally published 1957.)

Hidalgo-Downing, L. (2000) 'Negation in Discourse: A Text World Approach to Joseph Heller's *Catch-22*', *Language and Literature* 9(3): 215–39.

Jarausch, K. H., Arminger, G. and Thaller, M. (1985) *Quantitative Methoden in der Geschichtswissenschaft*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Kenny, A. (1992) 'Computers and the Humanities', Ninth British Library Research Lecture.

Kimbrough, R. (ed.) (1988) *Joseph Conrad: Heart of Darkness*, critical edn, 3rd edn. New York: Norton.

LaBrasca, R. (1988) 'Two visions of "The Horror!"', in R. Kimbrough (ed.) *Joseph Conrad: Heart of Darkness*, critical edn, 3rd edn, pp. 288–93. New York: Norton. (Originally published 1979.)

Leavis, F. R. (1962) *The Great Tradition*. London: Chatto & Windus.

Lothe, J. (2000) *Narrative in Fiction and Film*. Oxford: Oxford University Press.

Lothe, J. (2001) 'Cumulative Intertextuality in *Heart of Darkness*', in G. Fincham and A. de Lange (eds) *Conrad at the Millennium*, pp. 177–96. New York: Columbia University Press.

Louw, B. (1993) 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies', in M. Baker, G. Francis and E. Tognini-Bonelli (eds) *Text and Technology*, pp. 157–76. Amsterdam: Benjamins.

McMahon, A. and McMahon, R. (2003) 'Finding Families: Quantitative Methods in Language Classification', *Transactions of the Philological Society* 101(1): 7–55.

Phillips, M. K. (1985) *Aspects of Text Structure*. Amsterdam: North Holland.

Sarvan, C. P. (1988) 'Racism and the *Heart of Darkness*' in R. Kimbrough (ed.) *Joseph Conrad: Heart of Darkness*, critical edn, 3rd edn, pp. 280–85. New York: Norton. (Originally published 1980.)

Scott, M. (1997) *WordSmith Tools* (Software). Oxford: Oxford University Press.

Semino, E. and Short, M. (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.

Senn, W. (1980) *Conrad's Narrative Voice*. Bern: Francke.

Sinclair, J. M. (1975) 'The Linguistic Basis of Style', in H. Ringbom (ed.) *Style and Text*, pp. 75–89. Stockholm: Sprakforlaget Skriptov AB & Abo Akademi.

Singh, F. B. (1988) 'The Colonialistic Bias of *Heart of Darkness*', in R. Kimbrough (ed.) *Joseph Conrad: Heart of Darkness*, critical edn, 3rd edn, pp. 268–80. New York: Norton. (Originally published 1978.)

Stampfl, B. (1991) 'Marlow's Rhetoric of (Self-)Deception in *Heart of Darkness*', *Modern Fiction Studies* 37: 183–96.

Stubbs, M. (2001) *Words and Phrases*. Oxford: Blackwell.

Stubbs, M. (2004) 'Conrad, Concordance, Collocation. Heart of Darkness or Light at the End of the Tunnel?' Third Sinclair Open Lecture. University of Birmingham.

Stubbs, M. and Barth, I. (2003) 'Using Recurrent Phrases as Text-type Discriminators', *Functions of Language* 10(1): 65–108.

Tabata, T. (1995) 'Narrative Style and the Frequencies of Very Common Words', *English Corpus Linguistics* 2: 91–109.

Watt, I. (1960) 'The First Paragraph of *The Ambassadors*: an Explication', *Essays in Criticism* 10: 250–74.

Watt, I. (1988) 'Impressionism and symbolism in *Heart of Darkness*', in R. Kimbrough (ed.) *Joseph Conrad: Heart of Darkness*', pp. 311–36. New York: Norton. (Originally published 1979.)

Werth, P. (1995) 'World Enough and Time', in P. Verdonk and J. J. Weber (eds) *Twentieth Century Fiction*, pp. 181–205. London: Routledge.

Youmans, G. (1990) 'Measuring Lexical Style and Competence: the Type-token Vocabulary Curve', *Style* 24(4): 584–99.

Youmans, G. (1991) 'A New Tool for Discourse Analysis: the Vocabulary Management Profile', *Language* 67(4): 763–89.

Address

Michael Stubbs, FB2 Anglistik, University of Trier, D-54286 Trier, Germany.
[email:stubbs@uni-trier.de]