A ROUGH GUIDE TO DOING CORPUS STYLISTICS

Tania M. G. Shepherd
(Universidade do Estado do Rio de Janeiro)

Tony Berber Sardinha
(Pontifícia Universidade Católica/SP)

ABSTRACT

This article has two main purposes. The first is to provide a short panorama of existing trends within computer-assisted stylistics. The second is to analyse a prize winning novel by English writer Julian Barnes, by resorting to the tenets and working tools of one of the newest branch of Stylistics, the so-called Corpus stylistics. To this end, the article starts by looking at various attempts at defining what style is and their implications to the definition of the discipline known as Stylistics. Then the paper presents recent work within the field of Corpus stylistics, as it describes the uses of computational tools as part of the stylistician tool kit. Finally, the paper provides a variety of ways with which a literary work may be approached digitally with a view to showing how computational tools can aid the stylistician in acts of interpretation.

Keywords: style; corpus stylistics; corpus-driven analysis; literature; Julian Barnes

Introduction or the trouble with 'style' and Stylistics

Style is a fuzzy concept. Style has been described as both pervasive and elusive because "most of us speak about it even lovingly, though few of us are willing to say precisely what it means" (ENKVIST, 1973, p. 11). Wales (2012) states that etymologically, style (*stylus*) was an instrument for writing, a kind of pen, and came to mean 'manner of writing' by metonymic change. The author adds that "the incisiveness of that instrument still reverberates, and that as stylisticians we remain

sh sharp to the weighty implications of any choice of features for the framing of different realities in literary text worlds and everyday life" (WALES, 2012, p. 11).

Hence, because of its etymological origin, style may be seen as the 'manner' in which something is done or someone does it. Seen this way, style can be understood uncontroversially when applied to language studies, as the way in which language is used in a given context, by a given person, for a given purpose (cf. LEECH; SHORT, 2007). In other words, style could be interpreted as the reflection of an author's personality or of his linguistic habits (e.g. the style of Machado de Assis); at other times it may be understood as the way 1anguage is used in a particular genre (e.g. the style of historical novels), period and school of writing (e.g. early eighteenth-century style), or some combination of these.

Seen from this dualistic perspective, a term used by Leech and Short (1981) to explain certain strands in stylistics, style is the form of a certain content which is adopted by a certain user, and thus anything which may express that which is particular , unusual or/ and deviant. This concept is similar to that of Biber and Conrad (2009, p. 144), i.e., who sees style (in fiction) as made up of deliberate choices by authors depending on how they want to convey a story. As a result any analysis of fiction must "cover characteristics of the imaginary world and choices of style whose functions are associated more with aesthetic preferences than the real-world."

There is yet a third interpretation for style, which Leech and Short (1981) call pluralistic. From a pluralistic perspective, any use of language is the result of choices at different levels from the overall linguistic system. Style is therefore relational and is seen in comparison to choices which could have been made, or rather by contrasting a set of real language choices made against the range of possible existing choices. Viewed from this angle, style may be confused with the Hallidayan notion of register in that, given a certain context of culture and situation, there will always be a linguistic choice made from other possible choices. This may seem to be an arbitrary and circular definition, or rather, before style is defined, it will have been delineated against an adequate 'norm'.

Whatever perspective is opted for, there always seems to be a mismatch between a 'text' and its 'style', however one defines either of the terms. Wales (2012) claims that in her attempt to write the

entry for 'style' in the 3rd edition of the *Dictionary of Stylistics*, her sub-headings for *style* increased exponentially, and yet she was unsure whether she had fully captured the essence of what style is. As an afterthought, she admits that she shares Carter's view of style as essentially dynamic – or rather, that there is styling, a process, rather than style, a product.

The difficulty to define what style is increases substantially when style is understood as the object of study of the discipline known as Stylistics. Stylistics has indeed been defined as the study of style (WALES, 2001, p. 372), as the study of the language of literature using empirical evidence and linguistic theory (WYNNE, 2005) and as a "method of textual interpretation in which primacy of place is assigned to *language*" (SIMPSON, 2004, p. 2). Carter and Stockwell (2012) state that the study of style for a long time oscillated between the critical practices derived from literary studies and the "rigour of descriptive analysis and a scientific concern for transparency and replicability in that description" and as a result Stylistics has suffered from a sort of split personality. They voiced their discomfort about the hybridism of the area by saying that

> As an academic discipline stylistics has tended to be seen, pretty much throughout the twentieth century, as neither one thing nor the other, or, possibly worse, as all things to all men and women, as sitting therefore uncomfortably on the bridge between the linguistic and the literary. Linguists have felt stylistics is too soft to be taken too seriously, tending to introduce irrelevant notions such as performance data and readerly interpretation; literature specialists, by contrast, have felt that stylistics is too mechanistic and reductive, saying nothing significant about historical context or aesthetic theory, eschewing evaluation for the most part in the interests of a naïve scientism and claiming too much for interpretations that were at best merely text-immanent. For one group, stylistics simply and reductively dissects its object; for the other, the object simply cannot be described in a scientifically replicable and transparent manner. (CARTER; STOCKWELL, 2012, p. xx)

Of course, work on literary texts has at times counted on the rigorous, replicable procedures of statistically-based compilation of evidence, known as Stylometry (see BURROWS, 1987, for an example). In this case, interpretation of textual features, namely, connectives, collocations, preferred syntactic structures, aims at attributing

authorship. According to Hoover (2003, p. 261- 262), whereas "stylistic analysis is more likely to be interested in large numbers of characteristics that together help to describe the styles of authors", for authorship attribution, "a small number of items may be sufficient to allow the reasonably confident attribution of a disputed text". This is done invariably in comparative terms in order to authenticate a text as having been written by author Z rather than Y, a task which may be pertinent to forensic linguistics as well.

Stylistics, however, is not generally known for replicable rigour because even though more than one theoretician has proposed a checklist of procedures (see SHORT, 1996 and STOCKWELL, 2002), the end product of a stylistic analysis is interpretation, which is, in turn, anchored on the reader-analyst.

If the end product of Stylistics is textual interpretation, then it could be claimed that there is an overlap between Stylistics and Criticism. However, nothing could be further away from the truth. Carter (2004) has made the difference between the two areas quite clear: whereas practical criticism implies an interpretive account of the text, stylisticians need to be able to make others see how the interpretive account has been reached.

So what is Stylistics? It may be defined as a field of empirical inquiry, in which the insights and concepts of linguistic theory are applied to analyse texts, both literary and non literary. Stylistics is therefore something one does, as one provides explications for how texts may be understood and interpreted by readers, mainly by resorting to linguistic insights as metalangue (CARTER, 2004). Doing stylistics implies resorting to one of the many models of linguistic analyses at our disposal in order to illuminate the process whereby a particular interpretation of text is formed, or new aspects of text revealed (see SHORT, 1995, p. 53).

A typical way to do stylistics is to apply the systems of categorisation and analysis of linguistics to poems and prose (fictional or otherwise), using theories related to, for example, phonetics, syntax and semantics (WYNNE, 2005). Thus, as language can be viewed from cognitive, psychological, feminist and discoursal perspectives, so can Stylistics, giving rise to subsets of the discipline which have been termed Cognitive Stylistics, Discourse Stylistics, Feminist Stylistics, Functional Stylistics, Pragmatic Stylistics and so on.

Corpus Stylistics

The latest linguistic turn within Stylistics is that of Corpus linguistics. Digital or digitalized corpora, especially those made up of literary texts have gradually become available either from the internet or from digital media. Thus, Corpus linguistics insights, using these data bases, have been incorporated in interpretative textual analysis.

Thus, in the last forty years Corpus linguistics has "spawned, or at least facilitated the exploration of , new theories of language-theories which draw their inspiration from attested language use" (McENERY; HARDIE, 2012, p. 1) and as a result has gained its own status as a *bona fide* branch of Linguistics (see SHEPHERD, 2009 for a lengthy discussion).

However, as Malbergh (2013, p. 1) remarks, "using computers to aid the analysis of literature does not seem an obvious choice". Only recently did Corpus Linguistics begin to be used in the treatment of literary texts, whereas there is a long tradition of applying quantitative and computational parameters to literary data. In addition, the analytical data which is generated by computers within Corpus linguistics has been seen by some as unfoundedly decontextualised. One starts from lexical items, or sets of lexical items and their lines of concordances, which are no more than excerpts from the original target text. In contrast, the analysis of literature carried out within the perspective of Stylistics presupposes close readings of a text with a view to explaining the same text. Thus, "a common complaint about corpus methods is that they avoid the qualitative analysis necessary for real understanding of stylistic effects" (McINTYRE, 2013, p. 410).

Admittedly, it might not always be possible to remove the apparent rigidity of certain corpus methods, but with the continuous development of new software, it is certainly possible to minimize it.

In fact there are, to date, three possible corpus analytical approaches to text: corpus-assisted analysis, (b) corpus-based analysis, and (c) corpus-driven analysis. These distinctions are not water- tight categorizations, but rather means of identifying common analytical practices, which may even be used in succession, in terms of trial and error approaches to a target text.

Corpus-assisted analyses may be carried out in order to check out a stylistician's intuition about the stylistic effects of a particular

target text. This way into the data neither requires the construction of specialist corpora (there are plenty of off-the-shelf corpora which can be used for corpus-assisted work), nor does it demand particular expertise in computing or corpus analytical techniques. Adolphs (2006) claims that when analysis is focused entirely on the target text in order to extrapolate information relating to that text alone, it may be seen as an *intratextual* analysis.

Corpus-based and corpus-driven analyses differ from corpus-assisted analysis, in that they treat the target analytical text (or texts) as a corpus in its own right. The analysis is then based on the comparison between the target corpus and a reference corpus. Adolphs (2006) calls this type of analysis *intertextual* analysis.

It is beyond the scope of this paper to provide a detailed list of the stylistic studies which have been corpus assisted, based or driven. Such a comprehensive coverage may be found in Ho (2011). However, Biber (2011, p. 15-16) has summarized the cross-fertilization between the two disciplines by arguing that

> ...a corpus provides the best way to represent a textual domain, and a corpus approach is the most powerful empirical approach for analyzing the patterns of language use in that domain. Such analyses are applicable in any-sub-discipline of linguistics that includes consideration of language use, including the study of lexical and grammatical variation, discourse patterns, spoken and written register variation, historical change, etc. I see the study of literature as no exception here.

Ways into a machine-readable corpus

The widely accepted tenets of Corpus Linguistics include its concern with meaning, and its insistence that the (single) word is not privileged, rather meaning is contained in lexical items (single words, compounds, multiword units, phrases, and even idioms). Finally the main belief behind all the work within Corpus linguistics is that frequency is an important parameter for detecting the meaning of a lexical item and for making general claims about the discourse (see TEUBERT, 2005, p. 5-6)

There are several ways of looking at a text and deciding – with the aid of a computer – what is significant, what is not. To begin with, most software packages offer an array of facilities to compute *n*-grams (also known as "clusters" or "lexical bundles"). *N*-grams are

items which tend to appear frequently together (*n* simply stands for any number) in particular genres or discourse types. By extracting *n*-grams from a corpus, one may have an insight into the corpus phraseology or terminology. In addition, by also extracting the same *n*-grams from a reference corpus – a larger compatible corpus which is used as a parameter of comparison with the study corpus - one can compare both lists of items and extract which items appear in the study corpus with a high enough frequency. These are then selected by the program to be keywords. Keyness – the name given to this phenomenon - is a useful indicator of style enabled by any software, as keywords may be indicative features of the study corpus, such as content lexical items, which in the case of narrative, may be plot explanatory. The work of Culpeper (2002) and Walker (2010) on keyness in *Romeo and Juliet* and *Talking it over*, is particular useful to establish characterization

Another way into a corpus involves tagging the corpus, in order to investigate a specific linguistic feature, or in order to find out about particular linguistic phenomena. The corpus may be tagged automatically at various levels of delicacy in terms of phonetic annotation (stressed syllables, intonation patterns and pauses), morphology (prefixes, suffixes, lemmas), lexis (part of speech tags) syntax, pragmatic annotation (types of speech acts) and semantics (see LEECH, 2004). In terms of annotation, the work of Semino and Short (2004) about speech and thought presentation, whereby a corpus was annotated in terms of direct and indirect speech and thought is particularly noteworthy.

A practical example of corpus stylistics: *The sense of an ending*

In this part of the article we will use Julian Barnes' 11th novel, *The sense of an ending* (henceforth, *TSE*), in order to show what computer tools and relevant strategies may highlight for a stylistics analysis.

First, we need to provide a summary of the plot of *TSE*. This is a short novel divided into two chapters of unequal length, both narrated by Tony Webster, a retired man in his sixties. Webster reminisces on his life as a young man, his friends (one in particular called Adrian), and a first love, called Veronica. He also tells us briefly that he got married, had a child, was divorced and now lives

an uneventful life by himself, as he had "wanted life not to bother (him) too much, and had succeeded – and how pitiful that was". The novel explores the unreliability of memory. At the beginning of the novel, the narrator claims that "what you end up remembering isn't always the same as what you have witnessed" (*TSE*, p. 3). Some one hundred pages later, Webster emphasizes his unreliability as a teller of his own life story:

> How often do we tell our own life story? How often do we adjust, embellish, make sly cuts? And the longer life goes on, the fewer are those around to challenge our account, to remind us that our life is not our life, merely the story we have told about our life. Told to others, but – mainly – to ourselves. (*TSE*, p.134)

The reader knows then and there that this narration may have been altered, embellished, cut and edited – and that the story that the narrator tells himself may not necessarily have been the story of his life.

As far as the title is concerned, it is borrowed from *The Sense of an Ending*, a famous book by literary critic Frank Kermode, which starts "It is not expected of literary critics as it is of poets that they should help us to make sense of our lives" (2000, p. 3), the assumption being that poets (as well as fiction writers) ought to do so. The sense of our lives, according to Kermode, is akin to understanding how it ends. The act of borrowing such a title from Kermode's book implies an invitation at intertextuality at many levels. Kermode has been said to be "among the first to theorize how ending constructs meaning in narrative ...he argues that it is only through the sense of an ending that readers are enabled to make sense of a narrative" (Ingersoll, 2007, p.15).

The title of Barnes' book is thus justifiably ambiguous. It provides an allusion to Kermode's 'endings', i.e., the end of the world and the end of one's life, but it is also a reference to the many plot endings a narrative may have, namely traditional closed endings, multiple endings, open endings, and all of these endings at the same time, depending on who tells/reads a story.

Having established what the title of *TSE* may signal, who the narrator may be, and the multiple purposes of his narration, we turn to the narrative itself. An initial point of entry into this narrative can be the placing of this book within the overall picture of Barnes' oeuvre. *TSE* is his 11th fictional work. The graph below (Figure 1) shows *TSE*, which was written in 2011, in contrast with the following novels: *Metroland* (1981), *Before she met me* (1982), *Flaubert's Parrot* (1984),

Staring at the sun (1986), *History of the world in 10 ½ chapters* (1989), *Talking in over* (1991), *Porcupine* (1992), *England, England* (1999), *Love, etc* (2000), *Arthur and George* (2005), *Lemon Table* (2006), *Nothing to be afraid of* (2009). Each line of the graph shows respectively, from top to bottom, the overall number of words (tokens), the total number of individual, non repeated words (types), and the standard ratio between the tokens and the types of each novel separately.

The number of tokens and types of each book may be obtained by feeding each individual book into text mining programs like Wordsmith Tools 5.0 (SCOTT, 2004) as .txt data. This may be done for instance, by first either scanning each novel with an OCR (Optical Character Recognition) software or by downloading them as eBooks and then transforming them into machine-readable texts. The standard token/type ratio (henceforth STTR) is often assumed to imply something about 'lexical density', although according to Scott (2004), STTR is a crude measure of this density. The STTR is calculated, according to Scott, by dividing the tokens and the types after every thousand words in each text file. The ratio is calculated for the first 1,000 running words, then calculated afresh for the next 1,000, and so on to the end of the corpus. An average is computed, which means that an average type/token ratio is obtained, based on consecutive 1,000-word chunks of text.

Instead of making an attempt at establishing the lexical density of *TSE* alone, we have calculated the STTR of every single fictional work by Julian Barnes in a time series: every data point in the series shows the STTR of each of his novels. The point here is to verify Barnes' writing style, or rather, how his STTR – or his lexical density – has varied thus far through his writing career.

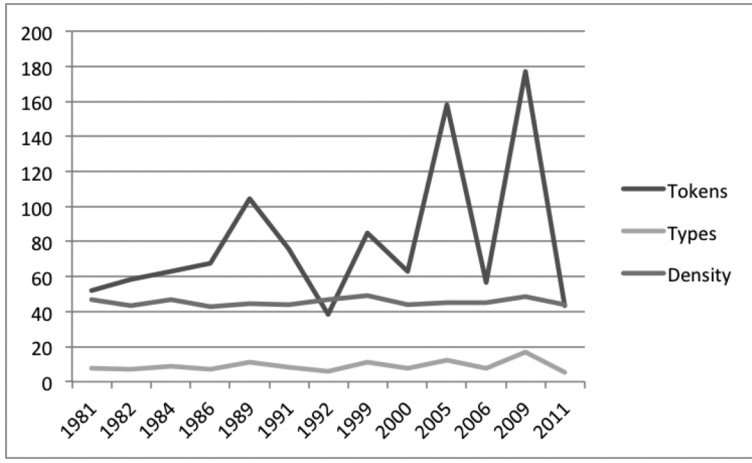

Figure 1: Lexical density of Barnes' novels

Figure 1 shows us the relative position of each novel in terms of number of words, number of individual words, and STTR. Although the number of words (top darker line) varies considerably from novel to novel, this is also matched by the number of types (middle-line). The density is thus kept nearly level from novel to novel, with the exception of the year 2009 when *Nothing to be frightened of* was written, Barnes' long personal memoirs and meditations about death.

A second way into the text, at a macro level, is the multi-dimensional approach to register variation, or MD Analysis for short. MD Analysis was originally developed by Biber (1988) for comparative analyses of spoken and written registers, but it has also been used in the analyses of literary texts. Unlike the previous approach, MD Analysis requires a high level of computer skills (see BERBER SARDINHA, 2013, for a detailed explanation).

The approach uses computational tools to identify a set of linguistic features which act as discriminators for a number of textual dimensions. A computational tool, i.e., the Biber grammatical tagger, is used to identify and tag a wide range of grammatical features in text, including word classes (e.g. nouns, modal verbs, prepositions),

syntactic constructions (e.g. relative clauses, conditional adverbial clauses, that-complement clauses introduced by nouns), semantic classes (e.g. activity verbs, likelihood adverbs), and lexical-grammatical classes (e.g. that-complement clauses introduced by mental verbs, to-complement clauses introduced by possibility adjectives), among many others in a text sample.

Multivariate statistical techniques allot these linguistic occurrences to five dimensions. Each 'dimension' carries a weight for each of the linguistic features identified, and as a last step, the program works out a corresponding level (positive or negative) for each dimension. In other words, by describing and quantifying the relative distributions of these features in a corpus, it is possible to interpret a text in terms of its functionality along the following five dimensions (from BIBER, 1988): 1. Involved versus informational production; 2. Narrative versus non-narrative concerns; 3. Elaborated versus situation-dependent Reference; 4. Overt expression of persuasion and 5. Abstract versus non-abstract style.

In the case of an author like Julian Barnes, the use of MD Analysis, which includes continuums of narrativity, informativity and abstractness, to cite just three of the poles by which textual data may be assessed, is particularly effective. Barnes is known to stretch the limits of what is accepted as narrative, both in terms of narration strategies and narrative organization proper (see SHEPHERD, 1993 and 1997).

To understand *TSE* in terms of the dimensions cited above, three texts were compiled and fed into the Biber tagger, namely, *TSE, Flaubert´s Parrot* (Barnes' most unusual narrative) in addition to all his factual writings, including *Nothing to Declare*. The program TagCount (a post-processor for the Biber Tagger) generated a chart containing percentages of frequency for every linguistic item for each and every work.

To calculate whether the corpus was more involved or more informative (Dimension 1), certain items were taken into consideration, namely pronouns, possessives, that-deletions, contractions, private verbs (think, believe, etc.), hedges, amplifiers, to cite a few (see BIBER, 1988, for a full list of the linguistic features associated with each dimension). The same procedure was carried out for each of the discriminators of the various Dimensions. The final chart obtained was the following:

	dim1	dim2	dim3	dim4	dim5
Flaubert´s Parrot	1.72	1.01	1.58	-1.01	1.83
The sense of an ending	11.19	2.05	-1.6	-0.94	2.93
Non Fiction	-4.5	-0.39	3.59	-1.83	1.92

Figure 2: MD Analysis of three samples of Julian Barnes' writing

Most work which starts from a computational standpoint is, of necessity, relational. Biber´s MD analysis is no exception. MD analysis starts from the assumption that no text is a true specimen of its kind. In other words, a text can be more involved, less narrative, more situationally dependent (the 'here and now' type of text), less persuasive and even so, more abstract, or any other combination of the poles of each dimension. In this way, any textual analysis escapes from the straight-jacket of classifying texts along the spoken-written continuum, or any one-dimensional way of characterizing texts or registers. Biber's own classification scores for each of the dimensions suggests that *TSE* may be viewed as a highly involved text, which displays certain characteristics of narrative, does not depend on the here and now for meaning, is not persuasive but has structures that express abstraction.

In terms of Dimension 1 (Involved versus informational production) and Dimension 2 (Narrative versus non-narrative concerns) the place occupied by Barnes' three sets of data is illustrated below. The place of the three corpora was calculated in relation to the various registers which were analysed by Biber and which were discussed in Berber Sardinha (2013).

Registers	Dim1
Telephone conversation	37,2
Face-to-face conversation	35,3
Personal letters	19,5
Spontaneous speeches	18,2
Interviews	17,1
SENSE OF AN ENDING	11,19
Romantic fiction	4,3
Prepared speeches	2,2
FLAUBERT'S PARROT	1,72
Adventure fiction	0
Mystery fiction	-0,2
General fiction	-0,8
Professional letters	-3,9
Broadcasts	-4,3
NON FICTION	-4,5
Science fiction	-6,1

Registers	Dim2
Romantic fiction	7,2
Mysteryfiction	6
General and science fiction	5,9
Adventure fiction	5,5
SENSE OF AN ENDING	2,05
Spontaneous speeches	1,3
FLAUBERT'S PARROT	1,01
Prepared speeches	0,7
Personal letter	0,3
NON FICTION	-0,39
Interviews	-1,1

Table 1: Dimensions 1 and 2 scores for *TSE*

In terms of Dimension 1 (involved versus detached concerns), the scores show that *TSE* is much closer to registers which are highly involved (and thus interpersonal), such as interviews, for instance, than to registers which pertain to fiction, such as romantic fiction. In terms of Dimension 2, which is narrative versus non- narrative concerns, *TSE* scores half way between the lowest scoring narrative fiction, which is adventure fiction, and spontaneous speeches, which are, on the whole, a register which is not characterized by narrative concerns.

It is, however, Dimension 5 which is the most revealing in terms of a writer such as Julian Barnes. Dimension 5 specifies the axis of abstractness versus concreteness dealt with in the registers analysed by Biber. The plotting of the three samples of Barnes's texts along this dimension produced the following diagram:

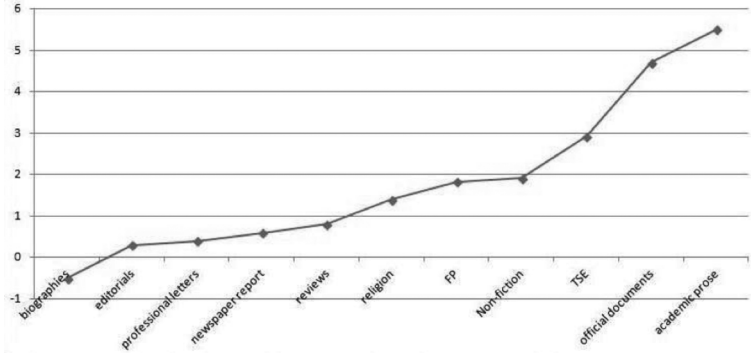

Figure 3: The comparative place of *TSE* within dimension 5

At one end of abstract concerns, there is academic prose, followed closely by official documents. The chart suggests that *TSE* may be viewed as high in the abstract continuum of dimension 5, next to official documents and Barnes' own non-fiction work. In fact, *TSE* ranks higher even in terms of abstract concerns than Barnes' most unorthodox novel, *Flaubert's Parrot*.

Thus, *TSE* may be described as a literary object whose concerns are highly involved for a novel – which may be explained by the fact that we are dealing with a first person narrative and a narrator who interacts with his readers. However, it is a narrative which is not entirely narrative and which deals with abstract themes. The MD analysis has helped the analyst obtain an X-ray of the novel as a whole, in relational terms, both in terms of everyday registers, and also in terms of other texts by the same writer. Having tackled the bigger picture, it is time we move on to a micro level analysis.

Another way into the corpus under study, with a view to analyzing it stylistically with the help of digital tools, is by counting its tokens (total number of running words) and types (different individual words used), and calculating the corpus keywords, or words with marked frequency. This is done by comparing the corpus list of words to the list of words of a reference corpus and comparing the frequency of each word in each corpus statistically. *WordSmith Tools* is the tool that originally made available keyword analysis, with its output formatted as a list. A program like Wmatrix, accessed at http://ucrel.lancs.ac.uk/wmatrix/ can create keyword in terms of 'clouds', a visual representation of the keywords, which enables the novice analyst

to appreciate the lexical items that are more prevalent in the corpus graphically. In the case of *TSE*, such keyword cloud is particularly effective in showing what is foregrounded in terms of themes.

The novel itself is divided into only two chapters, called no more than ONE and TWO, respectively. For analytical purposes, the two chapters were separated into two distinct sub-corpora, which were then fed into WMatrix. The program was asked to produce a Keyword cloud representing the two separate chapters. The ploy behind this procedure is to verify whether there are any changes in the number of times a certain word is repeated, which, in turn, places the word as a candidate for chapter keyword. The keyword cloud produced for Chapter One had the following keyword design:

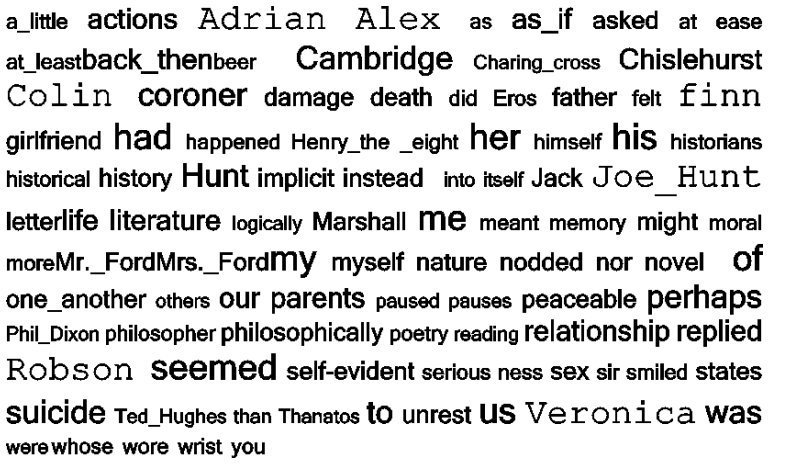

Figure 4 – Keyword cloud for chapter ONE of *TSE*.

In the case of chapter ONE, as expected, the names of the main characters appear in large size fonts, as the most important keywords in the cloud (Adrian, Alex, Colin, Robson, Veronica, Finn etc). Another set of words, belonging to the same semantic field, can easily be visualized as being one level lower in importance to the characters: 'coroner', 'suicide' and 'death', which may be said to set off the complicating actions in the plot. A few verbal processes 'asked', 'replied', 'states' and two behavioural processes ('nodded' and 'smiled') stand out, which are typical of narratives. However, most indicative

of the content of this chapter are three signals of the main character's tentative perceptions of his past: 'seemed', 'perhaps' and 'as if'. The protagonist of *TSE* says at the beginning of the narrative

> I need to return briefly to a few incidents that have grown into anecdotes , to some approximate memories which time has deformed into certainty. If I can't be sure of the actual events any more, I can at least be true to the impressions those facts left. That's the best I can manage. (*TSE* p.4)

And approximate memories they are. There are countless examples of imprecise memories signaled by 'seemed', 'perhaps' and 'as if'. For example, his perception of Veronica is based on impressions or fantasies:

> "When I was going out with her (Veronica), it always **seemed** that her actions were instinctive." (*TSE*, p. 35)

> "This ought to have made me feel accepted, but it **seemed** more as if they had grown tired of me." (*TSE*, p. 32)

> "Further, they (the books) **seemed** to be an organic continuation of her mind ..." (*TSE*, p.26)

When the protagonist justifies almost as a matter of fact why Veronica had allowed him to be intimate with her, he claims:

> "One evening, **perhaps a little drunk** , she let me put my hand down her knickers" (*TSE*, p. 36)

> "But I think I have an instinct for survival, for self-preservation. **Perhaps** this is what Veronica called cowardice and I called being peaceable." (*TSE* p. 45)

In terms of the narrator's use of 'seemed' and 'as if', they are fewer in number (and thus smaller in the keyword cloud) but equally key to the narrative. 'As if' has been identified as a style marker of imaginative prose fiction, occurring in fiction more often than in informative prose sections of textual data, according to Wikberg (1999) (*apud* MAHLHBERG, 2012). In fact, 'as if' implies an observation which attempts to compare two items, one of which is not immediately retrievable or is distant from the observer. 'As if' may be also seen as one of the 'little words' identified by Hunston (2012), in that it signals an interesting discourse pattern, that of a 'hypothetical-real' pattern, in which the 'real', i.e., the narrator's perceptions are contrasted to the 'hypothetical', i.e., to the 'as if' clause. By looking at the second

pair-part of the pattern, the 'hypothetical', one may have insights into the narrator's mindset: the protagonist's constant state of not knowing, of thinking that "Life seemed even more of a guessing game than usual" (*TSE,* p.41).

In chapter Two, the keywords are quite different in nature from those of chapter One. To the list of characters' names is added the word 'Tony', as a result of the narrator's reiteration of the end part of a torn letter "if Tony...". Margaret is also a keyword in the chapter. The first fifteen keywords in this chapter, in terms of keyness, are (according to the calculations obtained from WMatrix):

1. Veronica	6. had	11brother
2. Adrian	7. Margaret	12 of
3. life	8. email	13 Tony
4. my	9. diary	14memory
5. her	10 me	15 was

Figure 5 List of the 15 top keywords of chapter TWO of *TSE*

Because of the scope of this paper, the characters' names will not be focused upon, but the reader will be referred to Mahlberg's (2013) work on corpus and characterization. We will rather concentrate on the first lexical keyword of the list, namely 'life'.

The presence of 'life' as a keyword may confirm the abstract concerns of the novel. There are 95 instances of the word 'life' in chapter Two, a number which is comparatively high so as to make the item significant for exploration in this chapter.

The chapter starts "Later on in **life**, you expect a bit of rest, don't you ? You think you deserve it. But then you begin to understand that the reward of merit is not **life** 's business" (*TSE*, p. 65). This beginning contains the iteration of the item 'life' and prepares you for what the main protagonist fails to obtain, namely, a bit of rest. The concordance lines produced containing the word 'life' as the node word guide you through three possible meanings of the word 'life' in the story: the protagonist's difficulty of remembering life as

it was; the protagonist's realization of his wasted, purposeless life; the protagonist's own empty life in comparison to his friend Adrian's.

Well interspersed within chapter Two is the narrator's realization of not either a vaguely bland life or a life which could not be attested

> "Discovering, for example, that as the witnesses to your **life** diminish, there is less corroboration, and there is less certainty, as to what you are or have been". (*TSE*, p. 65)
>
> "We muddle along, we let life happen to us, we gradually build up a store of memories.
>
> The word resounded. Average at **life**; average at truth; morally average". (*TSE*, p. 97)
>
> "What did I know of **life**, I who had lived so carefully? Who had neither won nor lost, but just let **life** happen to him". (*TSE*, p. 155)

What is the purpose of life, the main character tries to answer?

> "If **life** is a wager, what form does the bet take?" (*TSE*, p. 93)
>
> "Sometimes I think the purpose of **life** is to reconcile us to its eventual loss by wearing us down, by proving, however long it takes, that **life** isn't all it's cracked up to be". (*TSE*, p. 115)
>
> "They were places where I always felt a sense of calm, odd as that may sound; also, a sense of purpose, perhaps the last proper purpose of my **life**". (*TSE*, p. 158)
>
> "Try as I could - which wasn't very hard - I rarely ended up fantasizing a markedly different **life** from the one that has been mine. I don't think this is complacency; it's more likely a lack of imagination, or ambition, or something . I suppose the truth is that, yes, I'm not odd enough not to have done the things I've ended up doing with my **life**". (*TSE*, p. 71)
>
> "I knew I couldn't change, or mend, anything now. You get towards the end of **life** - no, not **life** itself, but of something else: the end of any likelihood of change in that **life**. You are allowed a long moment of pause, time enough to ask the question: what else have I done wrong?" (*TSE*, p.163)

The narrator also compares himself to his friend Adrian, who had the clarity of taking control of his life:

> "I found myself comparing my **life** with Adrian's. ... The mental and physical courage of his suicide. "He took his own **life**" is the phrase; but Adrian also took charge of his own **life"**. (*TSE*, p.96)
>
> "I don't envy Adrian his death, but I envy him the clarity of his **life**. Not just because he saw, thought, felt and acted more clearly than the rest of us..." (*TSE*, p.114)
>
> "My philosopher friend, who gazed on **life** and decided that any responsible, thinking individual should have the right to reject this gift that had never been asked for..
>
> whose noble gesture reemphasized with each passing decade the compromise and littleness that most lives consist of. "Most lives": my **life"**. (*TSE*, p. 153)

Finally, having checked the novel's dimensions, inspected and compared the keywords for each individual chapter, we may resort to one of the last tools in the stylistician's digital toolkit, i.e., an inspection of the repeated multi-word units in the book, clusters of words (or n-grams) which appear together repeatedly in the corpus. Bigrams (two-word units) and less so three-grams are quite frequent in any corpus. Four-grams, however, are less frequent, but they do yield elements showing what the text is about (its terminology) and how it is structured (its phraseology). Five-grams are fewer and far apart. They require that clusters of five words are found verbatim with a minimum frequency, but if they do appear, especially in a literary text, they may be seen to foreground certain aspects of the fictional world being depicted, either a character's or the story teller's mannerism.

Just as an example, it was asked of the concordancing software to provide five-word clusters which were repeated at least three times throughout the novel. The list of five-grams produced consisted of 'but I didn't want to'; 'I don't expect you to' ; and 'you just don't get it', each cluster appearing three times in the novel. On inspection, the first cluster was found to be about the narrator's antagonistic feelings towards Veronica (I don't want to give her this pleasure/to think about her/to solve Veronica/to press Veronica). The second cluster 'I don't expect you to' refers to the narrator's repeated words to Veronica; I don't expect you to hand over Adrian´s diary/to reply to it (my letter)/ to think better of me. But the last five-gram 'you just don't get it', repeated three times as such and once as 'you **just** don't

get it' summarizes Veronica´s perceptions of the narrator: a man who has gone through life **not getting it**, not understanding life itself. A man whose life history was just "that certainty produced at the point where the imperfections of memory" met the inadequacies of documentation.

Conclusion

This study has been titled 'a rough guide' on purpose. Far from being a comprehensive stylistic analysis of a piece of writing, its original purpose was to show that Stylistics and computational tools can walk hand in hand to help stylisticians do the job of text interpretation by resorting to textual evidence. We have shown that, rather than pulling a piece of writing to pieces, we have attempted to turn the quantification of data into an aid of possible explication of the piece of writing. We have showed that computer aided tools can place a writer´s work within the universe of his other works and therefore help establish roughly how the writer tends to use his lexis. We have also showed that MD (multi dimensional) analysis is helpful in establishing the position of a particular piece of work in relation to other well described, every day registers – this is especially relevant in the case of writers who flout expected fiction narrative conventions. Finally, we have suggested that a lexical analysis of a text based on keyness, the extraction of concordancing lines for the keywords and a simple extraction of n-grams, may cast a light on the underlying (and often hiden) themes of a book.

To this digital toolkit, a number of other strategies could have been added, namely, quantifying, comparing and classifying characters' manners of gazing, body language lexis and speech habits; labeling and grouping the narrator's reported speech verbs; extracting significant collocates for action verbs, to cite a few of the existing computer aided approaches to Stylistics.

We feel it is time stylisticians and corpus linguists join forces and integrate each others' tools to produce solid, fresher, evidence-based analysis and explications of texts and move away from either Leavisite or interpretive criticism. Doing stylistics is not the prerogative of a chosen few with special sensibilities to unveil unique textual attributes. Doing stylistics by means of corpus analysis should be within the reach of anyone who firmly believes in the elucidating power of the lexis, who has access to a good text mining program

and who sees any level of linguistic organization (macro or micro) as a contributor to the text's overall style and meaning.

RESUMO

Este artigo tem duplo objetivo. Por um lado, apresenta um breve panorama de tendências no âmbito da Estilística assistida por instrumentos computacionais. Por outro, introduz estudo de um romance premiado do escritor inglês Julian Barnes, com base nos princípios e nos instrumentos analíticos de uma das mais novas áreas da Estilística, a chamada Estilística de corpus. Para tal, o artigo inicia pelo levantamento de diferentes concepções de estilo e suas implicações para a definição da disciplina conhecida como Estilística. Em seguida, o artigo apresenta pesquisa recente na área da Estilística de corpus, ao descrever as ferramentas computacionais que fazem parte da bagagem de trabalho do pesquisador na área da Estilística. Ao fim, o artigo introduz maneiras de abordar a obra literária digitalmente, com o objetivo de demonstrar como as ferramentas computacionais podem ajudar o pesquisador na área da Estilística em atos de interpretação.

PALAVRAS-CHAVE: estilo; estilística de corpus; análise com base em corpus; literatura; Julian Barnes.

REFERENCES

ADOLPHS, S. *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*. Abingdon: Routledge, 2006.

_____.; CARTER, R. Point of view and semantic prosodies in Virginia Woolf's *To the Lighthouse*. *Poetica*, v. 58, p. 7-20, 2002.

BERBER SARDINHA, T. Variação entre registros da internet. In: SHEPHERD, T.M.G.; SALIES, T.M.G. (orgs.) *Linguística da Internet*. São Paulo: Contexto, 2013, p. 55-76.

BIBER, D. Corpus Linguistics and the study of Literature: back to the future?. *The Scientific Study of Literature*. v.1, n.11, p. 15-23, 2011.

_____. *Variation Across Speech and Writing*. Cambridge: CUP, 1988.

_____.; CONRAD, S. *Register, Genre, and Style*. Cambridge: CUP, 2009.

BURROWS, J. F. *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon, 1987.

CARTER, R. *Language and creativity: The art of common talk*. London: Routledge, 2004.

_____; STOCKWELL, P. A celebration of style: Retrospect and prospect. *Language and Literature*, v. 21, p.9-11, Feb. 2012.

CULPEPER, J. Computers, Language and Characterisation: An Analysis of Six Characters in Romeo and Juliet'. In.: MELANDER-MARTTALA, U. et al. (eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium, Association Suedoise de Linguistique Appliquee (ASLA)*, 15. Universitetstryckeriet: Uppsala, 11-30, 2002.

ENKVIST, N.E. *Linguistic Stylistics*.The Hague: Mouton de Gruyter, 1973.

HO, Y. *Corpus Stylistics in Principles and Practice: A stylistic exploration of John Fowles´s The Magus*. London: Continuum, 2011.

HOOVER, D. Frequent collocations and authorial style. *Literary and Linguistic Computing,* v. 18, n. , 261-286, 2003.

HUNSTON, S. *Começando com as palavras pequenas*. In: SHEPHERD, T.M.G.. et al. (orgs.) *Caminhos da Linguística de Corpus*. Campinas, Mercado de Letras, 2012. p.31-64

INGERSOLL, E. G. *Waiting for the end: gender and ending in the contemporary novel*. Madison, New Jersey: Fairleigh Dickinson Univ Pr, 2007.

LEECH, G. Adding Linguistic Annotation. In: *M. Wynne* (ed.) *Developing Linguistic Corpora: a Guide to Good Practice,* Oxford: Oxbow Books: 17-29, 2005. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 03.06.2013].

_____; SHORT, M. H. *Style in fiction: a linguistic introduction to English fictional prose*. London: Longman, 2007. New revised edition.

_____; _____. *Style in fiction: a linguistic introduction to English fictional prose*. London: Longman, 1981.

MAHLBERG, M. *Corpus Stylistics and Dickens's Fiction*. London: Routledge, 2013.

McENERY, T.; HARDIE, A. *Corpus Linguistics*. Cambridge: Cambridge University Press, 2012.

McINTYRE, D. Corpora and Literature. In.: CHAPELLE, C. (ed.) *The Encyclopedia of Applied Linguistics*. London: Blackwell Publishing, 2013.

SCOTT, M. *WordSmith Tools*. 5.0. Manual. Oxford: Oxford University Press, 2004.

SEMINO, E.; SHORT.M. H. *Corpus Stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge, 2004.

SHEPHERD, T.M.G. O Estatuto da Linguística de Corpus: Metodologia ou área da Linguística? *Matraga,* Rio de Janeiro, v.16, n.24, jan./jun, 2009

_____.Towards a description of atypical narratives: a study of the underlying organization of Flaubert´s parrot. *Language and Discourse*. V. 5. p. 7 1-9, 1997.

_____. *A linguistic approach to the description of repeated elements in fringe narratives: principles of organization of prose and filmic text*. Unpublished PhD thesis. University of Birmingham, 1993.

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

_____. 'Understanding conversational undercurrents in *The Ebony Tower* by John Fowles'. In.: VERDONK, P.; WEBBER, J. (eds.) *Twentieth Century Fiction: From Text to Context*. Routledge: London, 1995, P. 45-62.

SIMPSON, P. *Stylistics: A Resource Book for Students*. Routledge: London, 2004.

STOCKWELL, P. A Stylistics Manifesto. In Szilvia Czabi and Judit Kerkowitz (eds.) *Textual Secrets: the message of the medium*. Martinvasar, Hungary: Akademia Nyomda, 2002, p. 742-758

TEUBERT, W. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1): 1–13, 2005.

WALES, K. *A Dictionary of Stylistics*. Harlow: Longman, 2001.

WALES, K. A celebration of style: Retrospect and prospect. *Language and Literature* 21(1) 9–11, 2012.

WALKER, B. Wmatrix, key-concepts and the narrators in Julian Barnes' Talking It Over. In Busse, B. and McIntyre, D. (eds.) *Language and Style*, 2010 , p. 364-387.

WYNNE, M (ed.). 2005. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 2005. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 03.06.2013].

Recebido em 30 de maio.

Aprovado em 15 de junho.