The potentialities of corpus-based techniques for analyzing literature
Khalid Shakir Hussein
Thi-Qar University, College of Education, English Department, Iraq
Email: khalidshakir74@gmail.com

Abstract:
This paper presents an attempt to explore the analytical potential of five corpus-based techniques: concordances, frequency lists, keyword lists, collocate lists, and dispersion plots. The basic question addressed is related to the contribution that these techniques make to gain more objective and insightful knowledge of the way literary meanings are encoded and of the way the literary language is organized. Three sizable English novels (Joyc's Ulysses, Woolf's The Waves, and Faulkner's As I Lay Dying) are laid to corpus linguistic analysis. It is only by virtue of corpus-based techniques that huge amounts of literary data are analyzable. Otherwise, the data will keep on to be not more than several lines of poetry or short excerpts of narrative. The corpus-based techniques presented throughout this paper contribute more or less to a sort of rigorous interpretation of literary texts far from the intuitive approaches usually utilized in traditional stylistics.

Keywords
Corpus Stylistics, Collocate Lists analysis, Dispersion Plots Analysis, Keyword Lists Analysis, Collocate Lists Analysis, Frequency Lists analysis.

Citation
Introduction

The analytic potential of certain corpus techniques might sound great in gaining some insightful knowledge that can be used in generating a sort of rigorous understanding of literary meanings and the way language is organized in a literary text. Ever since linguists started using corpora they have been thinking hard about conducting linguistic analysis of various types of electronically stored data (everyday conversations, newspaper editorials, emails, etc.).

Electronically stored literary data (texts) are no exception. Such corpus constitutes a rich repertoire of a highly complex use of language. It is the first time that huge quantities of literary language is completely reachable and open to systematic analysis and detailed description. No more intuitive analyses. Intuition is most definitely unworkable under the unprecedented sizable amounts of corpora. What is analyzed is not an excerpt of a novel, or a few lines of a poem but a corpus that holds for 469,720 tokens, which is the amount of data comprised throughout this paper (see Fischer-Starcke, 2010).

Theoretical underpinning: Empirical Corpus-based methods

Under the pressure of the ever-growing types of corpora, corpus linguists were completely motivated to impose some methodological mould on what looked like a forest of wild and out-of-control bushes of corpora. However, some sort of agreement should be achieved on the basic methodologies used in collecting and analyzing data so that it would be possible to compare and draw conclusions out of such large bodies of data crunched within different types of corpora.

The methodological issue represents one of the salient and pervasive concerns that stimulated the real motive behind the birth of corpus linguistics. The introspective judgments of native speakers used to be the appropriate source of data for the linguistic analysis. This view has been held for a long time under the overwhelming effect of Chomskyan linguistics. Thus, the rationalist methodology flourished and became the most reliable framework used to validate the use of the linguist's introspection as the only acceptable source of data (Sampson, 1980:150-51). Nevertheless, empirically oriented linguists insisted that the linguist's introspection should not be treated as authoritative (ibid:151). Such an introspection might be useful, as Chomsky amazingly proved its rich potentialities, but it is extremely necessary for this introspection to be verified by hard evidence drawn from some representative corpora.

This growing methodological debate and awareness, coincided with the emergence of unprecedented large bodies of electronic data, led linguists dramatically to conduct large-scale corpus-based investigations. An investigation of this kind made it impossible to deny the crucial role corpus-linguistic methods
The potentialities of Corpus-based techniques for analysing literature

and tools play in any empirical study of language-use (Pezik, 2011:447). Approaches to language whether theoretical or practical should use computerized corpora so that they can be placed on a firm empirical foundation. In the remaining part of this paper the researcher will figure out the exploratory potential of five tools and methods used to process various literary corpora.

METHODOLOGY
The methodology used in this paper involves building a corpus for each English novel to be studied by using the readily available electronic data (machine-readable corpus). Three huge novels are considered: James Joyce’s Ulysses, Virginia Woolf’s The Waves, William Faulkner’s As I Lay Dying. The texts have been selected for the heated literary debates they stir among literature scholars. Here comes the role of WordSmith tools to see how far they can contribute to a more rigorous exploration of the complex patterns revealed in the language utilized in these three novels. After establishing each corpus, the researcher will use (WordSmith Tools version 0.5) to go through the various analytic processes such package of programs supports.

WordSmith Tools Version (5.0) represents one good and up-to-date example showing how friendly the programs could be in processing linguistic data. As Scott (2010: 2), the programmer, puts it : "WordSmith Tools is an integrated suite of programs for looking at how words behave in texts." This suite of programs comprises three tools: the WordList tool (lets you see a list of all the words or word-clusters in a text, set out in alphabetical or frequency order); the concordancer, Concord, (gives you a chance to see any word or phrase in context); and Keywords with which you can find the keywords in a text) (ibid.).

The tools are widely used by Oxford University Press in working out some modern dictionaries, by language teachers and students, and particularly by researchers interested in exploring linguistic patterns of different languages. These three tools will be used throughout this paper in addition to some minor but descriptively effective techniques which might be looked at as a byproduct analytic tools.

Needless to say, all the digital data will be transcribed into plain text format before being processed by WordSmith Tools.

RESULTS AND DISCUSSIONS
CONCORDANCES
No one works on any area in corpus linguistics would proceed without coming across the term concordance. Concordance analysis is probably one of the most popular corpus analytic tools that should be undertaken by any researcher interested in corpus-based studies. There is an incredible heap of definitions
offered by corpus linguists for what a concordance refers to, however, the one suggested by Sinclair (1991:32) might touch the core of concordance form and function:

A concordance is a collection of the occurrences of a word-form, each in its own textual environment. In its simplest form it is an index. Each word-form is indexed and a reference is given to the place of occurrence in a text.

*Word-form* might be simply replaced by *lemma*—"a set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling" (Francis & Kucera, 1982:1)—a matter which complicates the searching process since every word-form (singular, plural, gerund, etc.) should be searched independently.

The expected output of a concordance is a list of all the occurrences of a word-form in a particular corpus, together with its context in which it occurs—usually a few words to the left and right of the *search word*. This is why concordance programs are also referred to as KWIC (Key Words In Context) (Scott, 2010: 147).

Table (1) below shows a printout for a concordance in KWIC format. In this example the search word *death*, as it occurs in Woolf's novel *The Waves*, is presented at the center of a fixed context of words or characters. KWIC format here is very helpful in finding out the kind of grammatical structures and set phrases which co-occur with the search word. The printout in Table (1) represents a simple concordance list obtained by using a modern concordancer (*WordSmith Tools version 5*- Scott, 2010).

**Table (1) KWIC Concordance List of death in Woolf's The Waves**
The potentialities of Corpus-based techniques for analysing literature

It is quite evident that *death* tends to appear in strange comparative structures: (. . . is more truly *death* than the *death* of friends.), (. . ., than the *death* of youth.). Besides, there is an emphasis on the religious context that can explain the way Woolf discusses *death* throughout her novel: (. . .the valley of death . . .). Further more, the phrases that co-occur with *death* suggest Woolf’s tendency to use this word in prepositional phrases that indicate extreme quantifiers: (. . .he wears Mrs Ramsay to *death* . . .), (. . .must be bored to *death* . . .), (. . .resist tyranny to *death* . . .).

However, it is possible to display the search word in a number of ways. Each way has its own value under certain research considerations. The concordance extract in Table (2) displays an alternative way of viewing the data in which the search word *die* is shown with all its lexical forms or *lemma*. In this Table the concordancing software is not limited by displaying a single word-form as it is the case in Table (1) but it rather extends its possibilities to encompass concordances of the full word-forms or search string (*die**) that will give (die, dies, dying, died).

Die lemma shown in Table (2) presents *death* as a node word in concordance lines such as: (. . . it dies away . . .), (. . . and dies away), (. . . as the interest of the story died away in them . . .), (. . .the sound die on her ear. . .), (. . .as the resonance died. . .). Beside the literal meaning of *die*, the contextual structures of these lines collocate with one specific metaphorical meaning of *die* (to end gradually). This might highlight the significance of Woolf's stylistic choice to use *die lemma* in such a way rather than making some other choices to explain the same meaning.

Table (2) Concordance Sample for *die* Lemma in Woolf’s *The Waves*
Concordance programs are still highly productive analytic methods that make it possible to bring altogether the examples of a particular linguistic item available within the original context.

FREQUENCY LISTS
The production of frequency lists, together with the generation of concordances, constitute two core corpus-processing techniques (Evison, 2010: 122). To produce a frequency list for a particular corpus is to make a specialized software process all the items in the corpus establishing a basic statistics concerned with the total number of tokens and the number of types distributed across the totality of these tokens (ibid:124). The frequency-count of this type is of a great help in calculating the type/token ratio of a corpus. After processing the whole body of data the software displays the frequency list in one of two ways. First, the frequency list can be given according to the rank order of frequency-ranging from the commonest ones to those less common which might even occur only once in the whole body of the corpus. This type of display is commonly called raw data as Table (3) below shows.

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>THE</td>
<td>298</td>
<td>5.924453259</td>
</tr>
<tr>
<td>2</td>
<td>AND</td>
<td>197</td>
<td>3.916501045</td>
</tr>
<tr>
<td>3</td>
<td>IT</td>
<td>153</td>
<td>3.041749477</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>136</td>
<td>2.703777313</td>
</tr>
<tr>
<td>5</td>
<td>HE</td>
<td>107</td>
<td>2.127236605</td>
</tr>
<tr>
<td>6</td>
<td>I</td>
<td>98</td>
<td>1.948310137</td>
</tr>
<tr>
<td>7</td>
<td>TO</td>
<td>97</td>
<td>1.928429365</td>
</tr>
<tr>
<td>8</td>
<td>SAYS</td>
<td>75</td>
<td>1.4910537</td>
</tr>
<tr>
<td>9</td>
<td>IN</td>
<td>67</td>
<td>1.332008004</td>
</tr>
<tr>
<td>10</td>
<td>IS</td>
<td>61</td>
<td>1.212723613</td>
</tr>
<tr>
<td>11</td>
<td>OF</td>
<td>61</td>
<td>1.212723613</td>
</tr>
<tr>
<td>12</td>
<td>ON</td>
<td>58</td>
<td>1.153081536</td>
</tr>
<tr>
<td>13</td>
<td>YOU</td>
<td>52</td>
<td>1.033797264</td>
</tr>
<tr>
<td>14</td>
<td>WAS</td>
<td>42</td>
<td>0.834990084</td>
</tr>
<tr>
<td>15</td>
<td>HIS</td>
<td>41</td>
<td>0.815109372</td>
</tr>
<tr>
<td>16</td>
<td>THAT</td>
<td>39</td>
<td>0.775347888</td>
</tr>
<tr>
<td>17</td>
<td>UP</td>
<td>38</td>
<td>0.755467176</td>
</tr>
<tr>
<td>18</td>
<td>WITH</td>
<td>38</td>
<td>0.755467176</td>
</tr>
<tr>
<td>19</td>
<td>CASH</td>
<td>32</td>
<td>0.636182904</td>
</tr>
<tr>
<td>20</td>
<td>HIM</td>
<td>32</td>
<td>0.636182904</td>
</tr>
</tbody>
</table>

Table (3) Frequency List Extracted for the Top (20) Commonest Words (based on Faulkner's As I Lay Dying)
The potentialities of Corpus-based techniques for analysing literature

Table (3) above shows the beginning of a rank order in the frequency list (N) for a relatively small corpus with a size of (5,000) words sampled from Faulkner's novel (As I Lay Dying). A word and a token are considered to be the same thing in this Table. What is displayed in Table (3) is not only the rank order (N) and the raw frequency (number of occurrences) of each token, but even the percentage of every token in relation to the corpus totality. Frequencies are sometimes given percentages or proportions so that comparisons between corpora of different sizes can be made (Baker et al, 2006: 75).

An alternative way of displaying the final counts consists in listing the tokens frequency according to the alphabetical order. Table (4) outputs the final counts in this way giving a different picture of the same distribution of frequencies beginning from rank (39) to (58).

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>AS</td>
<td>21</td>
<td>0.416501403</td>
</tr>
<tr>
<td>40</td>
<td>BACK</td>
<td>15</td>
<td>0.297500998</td>
</tr>
<tr>
<td>41</td>
<td>BY</td>
<td>16</td>
<td>0.317334384</td>
</tr>
<tr>
<td>42</td>
<td>BUT</td>
<td>20</td>
<td>0.396667987</td>
</tr>
<tr>
<td>43</td>
<td>COULD</td>
<td>22</td>
<td>0.436334789</td>
</tr>
<tr>
<td>44</td>
<td>DOWN</td>
<td>15</td>
<td>0.297500998</td>
</tr>
<tr>
<td>45</td>
<td>FROM</td>
<td>19</td>
<td>0.376834601</td>
</tr>
<tr>
<td>46</td>
<td>GET</td>
<td>19</td>
<td>0.376834601</td>
</tr>
<tr>
<td>47</td>
<td>HAD</td>
<td>20</td>
<td>0.396667987</td>
</tr>
<tr>
<td>48</td>
<td>HORSE</td>
<td>16</td>
<td>0.317334384</td>
</tr>
<tr>
<td>49</td>
<td>IF</td>
<td>20</td>
<td>0.396667987</td>
</tr>
<tr>
<td>50</td>
<td>INTO</td>
<td>20</td>
<td>0.396667987</td>
</tr>
<tr>
<td>51</td>
<td>ONE</td>
<td>20</td>
<td>0.396667987</td>
</tr>
<tr>
<td>52</td>
<td>PA</td>
<td>18</td>
<td>0.357001185</td>
</tr>
<tr>
<td>53</td>
<td>SAID</td>
<td>19</td>
<td>0.376834601</td>
</tr>
<tr>
<td>54</td>
<td>SAY</td>
<td>21</td>
<td>0.416501403</td>
</tr>
<tr>
<td>55</td>
<td>SEE</td>
<td>18</td>
<td>0.357001185</td>
</tr>
<tr>
<td>56</td>
<td>THEN</td>
<td>15</td>
<td>0.297500998</td>
</tr>
<tr>
<td>57</td>
<td>TIME</td>
<td>18</td>
<td>0.357001185</td>
</tr>
<tr>
<td>58</td>
<td>WILL</td>
<td>20</td>
<td>0.396667987</td>
</tr>
</tbody>
</table>

Table (4) Frequency List Extracted from the First (60) Commonest Words in Alphabetic Order (based on Faulkner’s As I Lay Dying)

The usefulness of frequency lists lies in characterizing certain universal properties of texts, or even of languages in general (Scott, 2010:148). By comparing frequency lists of two or more corpora, corpus linguists can figure out
what kind of words make up the most frequent vocabulary items and how this can be related to the text-type or genre (ibid.). Of course, one of the most important benefits frequency lists can bring about is providing lexicographers with exceptionally useful documents about the commonest words used by speakers of a particular language (Halliday, 2004: 17).

Nevertheless, what holds my interest is the very particular relationship between the frequency counts of a particular lexical item and the latter's significance in a literary corpus. This paper assumes a sort of correlation between frequency and the stylistic significance of a particular lexical item or any other linguistic feature. To discuss the writing style of a particular text is to pin down the most frequent linguistic features that might be a distinctive indicator of that style. After all, style is defined, amongst other definitions, as recurrence (see Mukherjee, 2005). The significant content or structure of a specific linguistic item or feature might be justified by a recurrent tendency of using it. Therefore, the frequent uses of the, and, it in Faulkner's As I Lay Dying might suggest a distinctive indicator of Faulkner's style that could distinguish him from others. However, what matters in the thematic context is not the frequency counts of function words but the recurrent times of content words.

Table (4) highlights two content words as being the most frequent in Faulkner's As I Lay Dying: SEE and TIME. These two words might not haphazardly recurred throughout the novel. The narrator is dying and engaged in a vivid process of recalling. There is an intensive provoke of the passing time and the squeezed senses represented by seeing. These two words suggest the eagerness of the narrator to have a maximum sensual experience that could distil every possible minute to prolong the remaining time span.

However, one must be careful in overusing frequency lists in verifying significant literary themes. They should not be overburdened with too much explanatory power. As Biber et al (2004: 176) puts it:

We do not regard frequency data as explanatory. In fact we would argue for the opposite: frequency data identifies patterns that must be explained. The usefulness of frequency data (and corpus analysis generally) is that it identifies patterns of use that otherwise often go unnoticed by researchers.

Simply speaking, frequency lists do not explain themselves but they need to be coordinated with concordance-analyses so that they both might explain why certain particular words are used quite frequently. The context which concordancers provide is so crucial in highlighting the associations that might be held, for example, between the most frequent lexical items and the most frequent grammatical structures. Moreover, this type of data is still very
appealing in deciding the focal point of a text or comparing the most foregrounded lexical items in more than one text (Baker et al, 2006:76).

KEYWORD LISTS
It is not easy to find out what a keyword is. It might be a word which appears to occur in a particular corpus much more frequently than what is expected. Therefore, it could be extremely frequent in a very small number of texts in a particular corpus (ibid: 97). According to Scott (2010: 157), one can not have an idea about what is expected without using a particular reference. In WordSmith Tools version 5, the keyword program starts with word lists or frequency lists as described above. Two word lists must be made: one for the text or set of texts the researcher is interested in, and a second is made for some reference corpus which would be better if has a part-to-whole relation with the first corpus (Scott, 2010:159).

Accordingly, if a corpus, for example, involves (1,00) files with equal sizes, and a particular word occurs (75) times in one single file within the same corpus, this word could be a keyword. Any keyword program would classify words according to their keyness ranging from those with the highest keyness to those with the lowest. This kind of keyword lists that includes items significantly and extremely frequent is called Positive Keyword List (Evison,2010:127).

However, Negative Keyword Lists can be identified on the opposite side of the corpus margin. Negative keywords tend to appear significantly less often in the single file than in the reference one (ibid:128). Table (5) below shows the three most significantly infrequent words in a single file sampled from Joyce's novel Ulysses with a size of (5,104) tokens and compared with the size of the novel as a superset (269,850) tokens constituting a reference corpus which should always be larger.

<table>
<thead>
<tr>
<th>N</th>
<th>Keyword</th>
<th>Freq</th>
<th>%</th>
<th>RC. Freq</th>
<th>RC. %</th>
<th>Keyness</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HIS</td>
<td>27</td>
<td>0.5289965</td>
<td>3332</td>
<td>1.2347606</td>
<td>26.139691</td>
<td>3.146637</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>22</td>
<td>0.4310346</td>
<td>3009</td>
<td>1.1150636</td>
<td>27.842032</td>
<td>1.287107</td>
</tr>
<tr>
<td>3</td>
<td>HE</td>
<td>35</td>
<td>0.6857366</td>
<td>4233</td>
<td>1.5686492</td>
<td>32.127389</td>
<td>1.151418</td>
</tr>
</tbody>
</table>

Table (5) Negative Keyword List of Joyce's Ulysses
The Table above indicates clearly that the three most significantly outnumbered words are pronouns (his, I, he) and their unusually low frequencies in comparison with the reference corpus (RC) are reflected with negative figures (-26.13; -27.84; and -32.12 respectively). The negative keyness of these three pronouns uncovers one particular feature of Joyce's narrative tendencies. This simple statistics points explicitly at Joyce's orientation in avoiding the use of first person narration.

Nevertheless, different approaches may be taken towards specifying what a keyword is: Stubbs (1996: 166) discusses the possibility of assigning keyness to any word that could be looked at as focal in a corpus, but this focality has nothing to do with the statistical measures set by Scott (2010) above. Kennedy (1998: 251) goes even further, a keyword has nothing to do with neither the frequency counts nor focality, it is rather any word that is observed to be the subject of a concordance.

Whatever was the approach, keyword lists are especially useful for the analysis of various bodies of literary data comparing one corpus with another. Besides, they work as a yardstick to characterize different types of texts and genres.

Collocates constitute the words that surround a particular search word (Scott, 2010:121). The phenomenon of Collocation, as described by Firth (1957: 14), takes into account the very fact that certain words tend to occur in combination with each other within certain linguistic contexts. Therefore, a collocate is most definitely a word that exists in the surrounding environment of another word (Baker et al., 2006:37).

Collocate lists are dynamically interwoven with concordances. The latter produces the actual occurrence of a search word accompanied with its textual environment, so what is displayed is the search word centralized within its whatever context. The focus in the collocate lists, however, is not the search word but the company-words as distributed around (Scott, 2010:121-22). For example, WordSmith Tools version 5 provide researchers with an independent window within which collocational occurrences and their frequencies can be set in columns and rows. Table (6) shows the top seventeen collocates for the word death in Woolf's novel The Waves as a corpus, within a (-5) to (+5) span.

The Table below displays each word surrounding the search word death which the concordance was based on, besides the strength of the collocational relationship between every two words which is measured carefully. For example, the strength of the relationship between the and death is (0.007) which is weak though the scores the highest frequency of occurrence (31). This might sound
The potentialities of Corpus-based techniques for analysing literature

contradictory, but it is not. The reason behind this queer dichotomy of the weakness of collocational relationship versus the high frequency of occurrence is the very fact that the scored relationship between the and death is by no means an exclusive one: the occurs as a collocate word with many other words and not only with death.

Table (6) Collocate List for death in Woolf's The Waves

However, we might have many other lexical words with lower frequency but their strength of collocation is higher: for example, the word Tyranny in Table (6) has a stronger collocational relationship with death (0.160) but a lower frequency of occurrence (6). Nevertheless, it is still more illustrative collocate of death than the. It is quite crucial to be careful and accurate in observing both frequency and exclusivity of collocates. Once more this could be very indicative of Woolf’s negative perspective of death in The Waves.

Then, the table shows the total number of times a collocate occurs with the search word, and a total for the Left and Right of the search word which occupies the Center. The number of words to the left and right depends on the collocation horizon set by the researcher according to the questioning points he pursues (Scott, 2010:124). In Table (6), the set of individual frequencies to the left and to the right of the search word is (5), i.e. 5 words to the left and 5 words to the right and there is a central spot reserved for the search word itself.

Collocational analysis of this type is helpful for various reasons. Pezik (2011: 456) captures three basic reasons:
First of all, no description of language can be complete unless it does some justice to its phraseology. Language is highly idiomatic and lexicality (of which collocations are a most important aspect) has become a level of linguistic analysis . . . Collocations have also been found to be a revealing source of information about discourse-specific metaphors . . . Studying collocations is also crucial in identifying selectional restrictions and semantic prosodies, the latter of which can be defined as the \textit{attitudinal load} (italics mine) of certain lexical items.

The second and third reasons are extremely relevant in characterizing any literary corpus. However, Scott (2010: 129) simply points out the very use of making out such a kind of lists: they are made just to figure out where the collocates \textit{crop up} a lot. Table (6) evidently spots (L1) as the position where the collocates of the word \textit{death} crop up a lot (35 collocates). This designation might ascribe a sort of \textit{attitudinal load} to (L1) that should be taken into consideration as a possible distributional feature that might characterize the language used by Woolf.

\textbf{DISPERSION PLOTS}

This technique of analysis is a complementary part of concordance. It is dependent on the concordance and derived from its lists (ibid:130). An interesting and vivid visual representation can be obtained by dispersion plots showing how regular the distribution of a search word is over a particular corpus (Baker et al., 2006: 59-60). Normally, dispersion value is figured out mathematically using descriptive statistics to compute its distribution by the following formula (Pezik, 2011:454):

\[
D = 100 \times \left( 1 - \frac{V}{\sqrt{n} - 1} \right)
\]

\[n = \text{number of text files.}\]

\[V = \text{the result of dividing the standard deviation by the mean of frequencies of the search word in question.}\]

In WordSmith Tools version 5, dispersion plot helps the researcher determining where his search word occurs throughout his textual corpus. So that, he will have an exciting visual representation of the dispersion of the search word that enables him to spot where in the corpus his search word is mentioned most or he can even use this plot to have a better idea of "the words' evenness of distribution" (ibid.), i.e. how equally the word in question is distributed over the various parts of the corpus. The following two Tables (7) and (8) show the distributional behaviour (dispersion plots) of two words \textit{death} and \textit{life} in a relatively small corpus of Woolf's \textit{The Waves}. The two words are selected purposively: Woolf is said to be a suicidal character and haunted throughout \textit{The Waves} by \textit{death} with no recourse to the spectacular aspects of \textit{life}. Thus, the
The potentialities of Corpus-based techniques for analysing literature

researcher raises a simple question about the dispersion plots of the words death and life to figure out which one of them has a more intensive presence in the language used in The Waves.

Table (7) Dispersion Plot of death in Woolf’s The Waves

Table (8) Dispersion Plot of life in Woolf’s The Waves
The plots above show the following details:

*File*: source text file-name  
*Words*: number of words in the source text  
*Hits*: number of occurrences of the search word  
*Per 1,000*: how many times of occurrence per 1,000 words  
*Dispersion*: the plot dispersion value  
*Plot*: a plot showing where they cropped up, . . . (Scott, 2010: 129)

The dispersion plots of *death* and *life* show different frequencies or *hits* in the corpus (*death* = 66; and *life* = 213). What is more, *life* is more evenly dispersed across the novel, whereas *death* occurs in a fewer number of portions. Accordingly, the dispersion plot of *life* clearly suggests that *life* occurs more as a central theme in the debates involved in the novel than the word *death* since it seems to be a more focused subject of the corpus at its various parts as being sorted by number of words per 1,000.

**CONCLUSION**

The reason that makes the researcher focus on the five techniques tackled above lies not only in their relatively simple linguistic nature but even in the user-friendly software suites they are packed in. Thanks to Scott's WordSmith Tools (1996- ) which have become very powerful tools and quite promising methods that can be readily used in carrying out diversified types of linguistic data mining.

The paper ends up with one specific conclusion that it is not any more appropriate to be satisfied by the intuitive critical interpretations of a literary text. The corpus-based techniques tackled throughout this paper seem to fulfill the goal of increasing the objectivity of a literary analysis. By virtue of such techniques, the linguist finds himself in a position to utilize software that provides his analysis with neutral and impartial insights into the literary texts under investigation. This would most definitely help the linguist to escape the overwhelming impressions surrounding the reception of the literary texts and to bring out some invisible meanings that could be missed or unrecognized by the intuitions of traditional literary stylistics. It would not be feasible to make an exploratory survey of three sizable and controversial English novels without the electronic analytic potential that corpus linguistics most definitely has.

**FUTURE RESEARCH**

Needless to say, the techniques and tools surveyed so far are by no means assumed to be the only empirical techniques available for literary language-analysis within the traditions of corpus linguistics. Many other quite sophisticated and more competent methods have not been surveyed: *n-gram*
methods; Markov Methods; Hidden Markov Methods; Supervised and Unsupervised Learning Methods; Sparse data Method; etc. It is true that the vast majority of these methods rely heavily on employing quantitative statistical information drawn from corpora, however, they confine language to its algebraic properties. The use of such methods in future research assumes the researcher to be familiar with the probabilistic variants of the formal grammars in a way that sounds more mathematic than linguistic.

Moreover, there is as yet no agreement on the nature of corpora that such methods try to exploit. This last point might be understood in the light of the technological revolution that enabled researchers to work on unprecedented amount of large-scale corpora of various and highly technical resources. Therefore, it will be a serious challenge to verify the viability of more sophisticated techniques that might address more invisible and large-scale descriptions of the recurrent features that literary language conceals from the traditional stylistic approaches.

References


..........................................................