A CORPUS-BASED STUDY OF THE STYLE IN JANE AUSTEN'S NOVELS[1]

Raksangob Wijitsopon[2]

Abstract

While a corpus linguistic technique has been applied to various studies in text and discourse analysis, it has not been much adopted in stylistic analysis of literary texts. The present study, therefore, applies a corpus-driven approach to Jane Austen's six major novels, in order to see how well this new method works with literary texts, compared with what has been observed in previous studies of Jane Austen's language. It has been found that the corpus-driven approach can provide quite a few results that are useful in supporting and refining literary scholars' intuitive observations on the author's works. Some of the linguistic patterns derived from the comparative corpus-driven method have not been remarked on before in any previous studies and hence can serve as new textual evidence in the study of Jane Austen's writing style. Despite such great potential for the study of style in literary works, it is suggested that the analyst's knowledge and understanding of the text(s) under study is crucial in interpreting and evaluating those results because the corpus-driven approach to literary texts relies heavily on quantitative data.

Introduction

The past few decades have seen a remarkable prominence in the application of corpora in English applied linguistic research. This also includes the area of text and discourse analysis. The corpus-based technique allows text and discourse analysts to expand the size of their data, which in turn enables them to generalize their findings to a larger extent than before. On the theoretical side, it enables text and discourse researchers to show that patterns of co-occurrence among words in texts are associated with different meanings and uses in the communicative events (Sinclair 2004). For example, Biber and Conrad (1999) showed that conversation and academic writing are markedly different from each other through their frequency analysis of phrases found in corpora of the two text types. Recent work in critical discourse analysis, whose main concern lies in the relationship between discourse, ideology and power, also incorporates the use of corpora in its analytical practice. Moon and Caldas-Coulthard (2010), for instance, found from their analysis of a British newspaper corpus that women are frequently described in terms of their physical appearance, as shown by a high frequency of such adjectives as "beautiful", "pretty" and "lovely" used in collocation with references to women, whereas references to men are usually modified by adjectives related to importance, including "great", "key" and "main". This difference in the media's discursive practice, it is argued, reflects

[1] This study is sponsored by the TRF-CHE Research Grant for New Scholar and the Ratchadaphiseksomphot Endowment Fund of Chulalongkorn University (RES560530083-HS). I would like to express my gratitude to the Thailand Research Fund (TRF), the Commission on Higher Education (CHE), Ministry of Education, and Chulalongkorn University for their support.

[2] Assistant Professor, Department of English, Faculty of Arts, Chulalongkorn University

and simultaneously sustains the patriarchal ideological beliefs in British society.

While corpora have been adopted in the examination of a variety of text types, ranging from everyday conversation and newspaper reports to academic writing, little has been done on literary discourse. Given that our interpretation of a literary work relies particularly heavily on language in a text, it would be interesting to explore to what extent the use of corpora would enable us to investigate language in literary texts and its relationship with interpretative issues. To this end, I apply corpus techniques to the analysis of Jane Austen's six major novels. The research questions the present research addresses are:

(1) What lexical patterns are characteristic of Jane Austen's novels?

(2) What textual meanings do those patterns suggest?

(3) What are strengths and limitations of the corpus approach in the study of fictional prose?

In the sections that follow, I first give an overview of "corpus stylistics", the theoretical framework in which this research is grounded. Then, an outline of previous studies on Jane Austen's works and writing style is provided. This is followed by an account of the methodology adopted in this study, comprising explanations about corpus descriptive tools and data preparation. Then, the results of the study are reported, followed by a discussion on the strengths and limitations of the corpus-driven approach to literary texts, as observed from a corpus of Jane Austen's six major novels.

Corpus stylistics: The theoretical framework of the study

The term "corpus stylistics" has recently been used by scholars in stylistics (e.g. Short and Semino 2004) or in corpus linguistics (e.g. Mahlberg 2007), to refer to the practice of linguistic analysis of literary texts, making use of a collection of electronic texts, sampled to be maximally representative of a writer's works or a particular literary genre (see Biber 2011 for review of corpus stylistics research). While corpus linguistics is interested in describing normative uses of language, which can be inferred from repeated occurrences of linguistic patterns, stylistics pays particular attention to deviations from linguistic norms, which create particular textual meanings and aesthetic effects. Given the primary concern of each different discipline, corpus linguistics and stylistics appear to focus on opposing phenomena in the form-meaning relationship. However, in identifying deviant instances of language use in a literary text, stylisticians in effect draw on their observation or knowledge of what is normal in language use. Therefore, while linguistic deviation is central to stylistics, consideration of linguistic norms is in fact inherent to the practice of stylistic analysis.

It is through the concern about linguistic norms that stylistics and corpus linguistics come to converge. As Stubbs (2005: 21) notes, "[…] a text is a selection from the potential of the language […] Comparative corpus methods […] allow us to study how far texts consist of recurrent phrasal patterns which are widespread in the language as a whole." Corpus stylistics thus involves an explicit comparison between a corpus of texts under investigation and linguistic "norms",

represented by a corpus of the kind of texts that are "contextually related" (Enkvist 1973) to the text or group of texts under investigation.

Some corpus linguists (e.g. Tognini-Bonelli 2001, Mahlberg 2005) argue that a corpus-based approach can allow textual analysts to obtain quantitative data to test their hypothesis about textual features. On the other hand, an analyst can come to study the text without identifying what textual features are likely to mark the style of the work and keep their eyes open to the findings derived from a comparison between the text(s) in question and the appropriate reference corpus. These findings are then taken as textual patterns that are brought to analysts' attention and deserve further investigation in terms of the semantic and pragmatic roles they have in our interpretation of the text(s).

In the present study, I take this corpus-driven stance in that stylistic features of Jane Austen's novels are not first identified but, instead, the comparison is allowed to reveal lexical items and patterns that characterize Jane Austen's novels. These patterns will be analyzed qualitatively in order to see how they contribute to meanings of the texts under study.

Given that Jane Austen's six major novels are among canonical classic works in English literature, they were chosen to be the objects of study. As there have been numerous studies of Jane Austen's works, a brief note on her novels, particularly on her language use, is provided in the next section so that corpus-informed findings can be assessed and discussed in relation to intuitive observations made in previous studies of Jane Austen.

A brief review of previous studies of Jane Austen's novels

Jane Austen is one of the most renowned novelists in English literature. Her works usually present the story of a young lady who has some kind of limitations, including social standing, economic insecurity and even her own distorted understanding of the world. All of her heroines have to learn to overcome these limitations by developing self-understanding and sound judgment of people around them before they have a happy ending (McMaster 1996**;** Tanner 2007). Because Jane Austen's works involve the portrayal of the emotional, intellectual and spiritual growth of the female protagonists, they have often been criticized, according to Sherry (1966), for being lacking in physical action and full of socializing activities, such as neighbor or relative visits, picnics, and parties. Advocates of her novels, however, argue that Jane Austen is great in her realistic description and biting commentary on domestic life during the Regency period of England.

Although Jane Austen's works have been widely discussed in the study of English literature, little has been researched on the interplay between the author's linguistic choices and the aesthetic value of her novels. For example, literary critics often discuss the author's use of irony in her novels but rarely are the ironic statements explained as to how their ironic force is achieved (cf. e.g. Mudrick 1952). The few attempts that have been made to explain Jane Austen's language use are mostly concerned with her word choices. Page (1972), for instance, observes that Jane Austen's novels are full of abstract nouns. Booth (1991) also comments on the author's preference for abstract nouns,

suggesting that they are used to delineate reliable characters from comic or superficial ones, whose idiolects tend to be filled with concrete nouns. A more detailed account of Jane Austen's lexical choices is found in Stokes (1991), who argues that Jane Austen's novels contain four major groups of words, namely those related to (1) spirit, e.g. "vivacity" and "ardour", (2) manners, e.g. "civil" and "elegant", (3) intelligence, e.g. "accomplishment" and "discernment" and (4) temperament, e.g. "amiable" and "disposition", all of which are central to the development of the plots and themes of her novels, including those about judgment or the disparity between appearance and reality.

Given that Jane Austen's novels have been discussed widely in literary studies and that some observations have been made on her language use, though intuitively, an examination of the language in her six major novels using a new approach like corpus stylistics might be a boon to both English literary studies and corpus linguistics: findings from the corpus-driven approach can be compared with what has been said in the previous studies, which in turn would help us evaluate how well the approach works in relation to literary discourse; on the other hand, critics' careful observations of Jane Austen's writing can be validated or refined by findings from a systematic corpus-stylistic approach, which involves both the quantitative and qualitative analysis of all her six major novels.

Methodology

1. Descriptive tools

There are three corpus linguistic descriptive tools that are used in this study to explore stylistic features and their textual functions in Jane Austen's novels: "keyness", "collocation" and "cluster", each of which is explained in turn below.

a. Keyness

To answer the research questions stated above, the concept of "keyword" in corpus linguistics is drawn upon as a starting point in the analytical procedures. According to Scott and Tribble (2006), keywords are lexical items of significance to a text in question, because of their "unusual frequency in comparison with a reference corpus of some suitable kind". The "unusual frequency" here refers to both "unusually high" and "unusually low" frequency. For the purpose of the present study, only items with "unusually high" frequency are considered.

The phrase "unusually high" here suggests that keywords are not simply words of high frequency. In other words, keywords are not necessarily the most frequent words found in a text. Keywords are important to the text because they are used "unusually often" when compared with other texts. To illustrate, if we consider a corpus of Jane Austen's novels alone, the definite article "the" is found to be the most frequent word in Jane Austen's novels. However, can we immediately say that it marks Jane Austen's writing style? Given that "the" is an article, it is likely to be used very often in any piece of writing, not just in Jane Austen's novels. Therefore, we need to compare Jane Austen's works with other authors' so that we can see if she really used "the" significantly more than others. And when compared with other novelists' writing, the article does not turn up as a keyword in Jane Austen's works because it is also used frequently by other authors. On the

other hand, the word "very", whose frequency is lower than that of "the", appears in the keyword list of Jane Austen's novels. This means that Jane Austen used the word "very" significantly more often than other writers. Therefore, as Baker (2006) puts it, a keyword list gives a measure of saliency, not just frequency, of the lexical items in a text and hence can suggest further examination of their textual functions. This is the reason why keywords are fundamental to the corpus-stylistic analysis of Jane Austen's novels in the present study: through the Keyword function in Rayson's (2007) Wmatrix Tools (see below), lexical items that are characteristic of Jane Austen's novels are extracted, some of which will be further investigated in detail.

Based on the above principle, it can be seen that a keyword list is derived through a comparison between the text or corpus in question and reference corpora, with the size of each corpus and the frequencies of each word within them being cross-tabulated. Such statistical tests as the chi-square or log-likelihood tests are then employed to measure the degree of each word's significance. In this study, as far as a statistical test is concerned, the statistical measure log-likelihood is applied for all the comparisons. This is because, according to Leech et al. (2001), while many statistical tests rely on an assumption of normal distribution of data, which is often not the case with linguistic data such as word frequencies, the log-likelihood measure does not.

According to Scott and Tribble (2006), three kinds of words usually come out of a comparison as keywords: (1) proper nouns, (2) words that "human beings would recognise" as key, which tend to indicate a text's "aboutness", and (3) words that are not usually identified consciously by readers as key but nonetheless occur in significantly high frequencies and so can be indicators of the style of a text, rather than of its content.

(b) Collocation

After keywords are extracted, their significance in the six major novels needs to be explained. To this end, the concept of collocation, defined by Hoey (1991: 6-7) as "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context", is drawn upon. This definition emphasizes that collocation of a word is not just a random co-occurrence of words, e.g. "she + is", but the co-occurrence takes place in a text for some reason, as seen from the phrase "with greater than random probability in its (textual) context". For example, as shown by Stubbs (2001: 28), common collocations of the word "seek" include "help", "advice" and "support". An examination of the collocational patterns of a word in a text can therefore allow us to see the relationship between lexical items in a text, which in turn enables us to see the way words are used to create meanings in a text. To find out what keywords are used "with greater than random probability" in Jane Austen's novels, a computer-assisted extraction of collocates, through the statistical measure Mutual Information (MI)[3], is adopted.

(c) Cluster

In this study, while a collocation refers to an individual lexical item that is found

[3] Mutual Information is a concept in statistics, often adopted in lexicography to measure the strength of collocations (see, for example, Church and Hanks 1990)

through statistical measure to co-occur significantly with another lexical item, the term "cluster" is used to refer to a recurrent string of uninterrupted word forms, e.g. "you do not" and "I am sure that" (Scott 1999). Given these examples, clusters are thus phrasal constructions, which are combinations of lexis and grammar. According to Stubbs (2001), because clusters display both lexical and grammatical relationships among words, they play an important role in creating textual meanings. As far as the present study is concerned, the concept of "cluster" is of particular use in showing the way function words such as "could" and "must", which turn up in the keyword list of Jane Austen's novels, contribute to meanings in the author's works. As individual words, it is hard to explain systematically what semantic and pragmatic contributions the modal verbs "could" and "must" have to Jane Austen's major novels, because the verbs can be used for various purposes, including expressing probability, obligation, politeness or permission. However, if we consider their frequent clusters, e.g. "she could not say" and "you must have thought", we are able to see more clearly in what ways these function words are important to the style of Jane Austen's writing.

Clusters can be extracted automatically by the software WordSmith Tools (see below). To extract clusters from a text, we need to identify the length of a cluster. In the present study, I chose to look at three-word clusters, e.g. "I do not", because, after experiments on various lengths of clusters, this is the optimal length of clusters since a manageable size of data can be generated for detailed concordance analysis.

2. Data, reference corpora and software

To answer the research questions with the above three descriptive tools, three corpora were compiled, one being the main corpus data and the other two being reference corpora used for a comparison with the main corpus data:

(1) the main corpus data: a corpus of Jane Austen's six major novels (henceforth JA), with 729,488 tokens[4];

(2) the first reference corpus: a corpus of modern fictional texts from the British National Corpus provided by the Wmatrix software (henceforth "BNC");

(3) the second reference corpus: a corpus of British prose fiction published during the period 1780 – 1820 (henceforth 19CNov), with 2,027,118 tokens.

The latter two corpora serve as the reference corpora in this project, with which JA was compared so that key lexical items in JA can be extracted. BNC is meant to represent linguistic patterns in modern British fictional prose, with which modern readers are familiar and is therefore likely to affect readers' interpretation of Jane Austen's novels, while 19CNov is meant to represent the British English the author drew upon when writing her novels. In other words, BNC represents language on the receptive side while 19CNov represents language on the productive side of the texts under investigation. Thus, a comparison between JA and BNC would show how the language in Jane Austen's novels is

[4] The word "token" used here is a term in corpus linguistics, referring to an occurrence of any given word form.

different from the kind of English with which modern readers are familiar. A comparison with 19CNov, on the other hand, would show in what ways linguistic patterns in JA are different from those in the novels contemporary to JA. The lists of keywords obtained from the comparison of JA with the two different reference corpora are compared. The words that are found to occur on both lists are deemed true keywords that mark Jane Austen's writing style since they are found to be of statistical significance no matter what "linguistic norms" are considered, the tendency of present-day British English representing modern readers' language or that of 19th Century British English.

The software used for the corpora comparison in this project are WordSmith Tools, developed by Scott (1999) and Wmatrix, developed by Rayson (2007). Both are software tools for corpus analysis and text comparison. The former is an integrated suite of programmes that enable us to examine how words behave in texts. Wmatrix provides a web interface to the semantic and grammatical corpus annotation tools, i.e. USAS and CLAWS[5], respectively. Wmatrix users can upload their own corpus data to the system, so that it can be automatically annotated and viewed via the web browser. Wmatrix also extends the keyword method to key grammatical categories and key semantic fields, i.e. grammatical categories and semantic fields that are of significance to the text under investigation due to their unusual frequencies when compared with reference corpora.

An examination of the texts at three linguistic levels, i.e. lexical, grammatical and semantic, serves as a means of triangulation in the way we approach the texts, looking at different linguistic features of the texts, and also reduces the problem that may arise as a result of the focus on frequency and statistical value. That is, by focusing on keywords only, researchers may have to ignore words below the statistical cut-off point[6] even though they are actually closely related to those above the cut-off point. For example, as will be seen below, the word "very" is ranked second in the keyword list of JA. However, the author also used other degree adverbs, such as "so" and "really", but they do not turn up in the keyword list since their statistical value stands below the cut-off point set in the extraction. Nevertheless, they may be found as part of key grammatical categories if the "degree adverbs" category is designated as key. If not, we can infer that only the word "very" is of particular significance to Jane Austen's writing since it was used more significantly even than its close synonyms. In other words, the extraction of key semantic fields and key grammatical categories would help confirm the significance of the individual key words or shed light on the density of some words that may have been overlooked due to their relatively lower frequency as individual words.

[5] USAS stands for UCREL Semantic Analysis System and CLAWS for the Constituent Likelihood Automatic Word-tagging System

[6] The statistical cut-off point used here refers to the level at which the statistical value of a word is considered meaningful for data interpretation. For example, if a cut-off point is set at a statistical value of 0.05, words with values higher than 0.05 are considered important and should be chosen for detailed examination while those with values lower than 0.05 are not very important.

In this study, to extract key linguistic features, the cut-off point was set at the log-likelihood value of 200, which can be considered rather high. This is due to the fact that the two reference corpora are to a large extent different, one representing the language of present-day British English fictional prose and the other that of 19th century British English fiction. Consequently, lists of keywords derived from the comparison of JA with each reference corpus are likely to be remarkably different. Setting a high cut-off point means that the possibility of key linguistic features found on the lists occurring by chance is slim. Therefore, this would guarantee the keyness of the lexical items, grammatical categories and semantic fields that are found from the comparison. The items or categories that are found to turn up on both JA-BNC and JA-19CNov lists are considered true key items or categories in Jane Austen's novels and hence further examined through an analysis of their collocations and clusters.

Result

Based on the setting spelled out in the previous section, keywords, key grammatical categories and key semantic fields in Jane Austen's novels that are found from comparing JA with BNC and 19CNov are derived. However, as stated above, in this study only those that occur on both lists are considered significant style markers of Jane Austen's novels. The lists of keywords, semantic fields and grammatical categories in JA are presented in Tables 1-3 below, starting from the item with the highest degree of keyness.

Table 1: Keywords in Jane Austen's novels

Rank	Keyword
1.	be
2.	very
3.	not
4.	she
5.	her
6.	could
7.	every
8.	herself
9.	must
10.	such
11.	any
12.	been
13.	however
14.	sister
15.	feelings
16.	to
17.	have

Table 2: Key semantic fields in Jane Austen's novels

Rank	Key semantic field	Sample words in the semantic field[7]
1	degree boosters	very, so, much
2	likely	could, would, might
3	entire; maximum	all, any, every
4	thought, belief	think, felt, believe
5	kin	sister, father, mother
6	degree maximizers	most, perfectly, entirely
7	content	pleasure, glad, satisfied
8	strong obligation or necessity	must, should, obliged
9	social actions, states and processes	manner, visit, conduct
10	respected	respect, regard, esteem
11	expected	hope, expected, anticipated
12	like	dear, like, affection

Table 3: Key grammatical categories in Jane Austen's novels

Rank	Key grammatical categories	Sample words in the grammatical categories[8]
1	degree adverb	very, so, much
2	be – infinitive	be
3	noun of title	Mr, Miss, Mrs
4	modal auxiliary	could, would, might
5	determiner capable of pronominal function	such
6	have – infinitive	have
7	been	been
8	to	to
9	third-person singular objective personal pronoun	him, her

[7] The three most frequent words in each semantic field are given as sample words.
[8] If a grammatical category contains more than three items, e.g. the "degree adverb" group, the three most frequent words in each category are given as sample words. For those that contain three or fewer than, all of the items are put in the table.

Interpretation of overall findings

It is observed that a number of the findings relate to or overlap with one another, either in the same categories or across different linguistic groups. To illustrate, the keywords "she", "her" and "herself" are related in the sense that they are personal pronouns referring to women, or the keyword "very" is also part of the key grammatical category "degree adverb" and key semantic field "degree boosters", the latter of which also expresses a similar concept to the key semantic field "degree maximizers". Such correspondence suggests that these overlapping items or categories are especially characteristic of Jane Austen's novels since they remain on the lists, whether we look at individual lexical items, grammatical categories or semantic fields in the novels. Such overlapping items and categories can be put together into groups, resulting in a total of six groups of key linguistic features that mark the style of Jane Austen's novels. They are:

(1) words related to a high degree, which comprise the keywords "very", "every" and "any", the semantic fields "degree boosters", "entire, maximum" and "degree maximizers" and the grammatical category "degree adverb"

(2) modal auxiliary verbs, which comprise the keywords "could" and "must", the semantic fields "likely" and "strong obligation or necessity" and the key grammatical category "modal auxiliary"

(3) auxiliary BE and HAVE, which comprise the keywords "be", "been" and "have", the grammatical categories "be-infinitive", "have-infinitive" and "been"

(4) words related to women, which comprise the keywords "she", "her" and "herself" and the grammatical categories "noun of title" and "third-person singular objective personal pronoun"

(5) words related to family relationships, which comprise the keyword "sister" and the key semantic field "kin"

(6) words related to internal states of mind, which comprise the keyword "feelings" and the semantic fields "thought, belief", "content", "respected", "expected" and "like"

Applying Scott and Tribble's (2006) categorization of keywords (see above), these six groups of key linguistic features in Jane Austen's novels can be divided into two main groups. The first group contains lexical items that are likely to be identified through human observation and suggest the content of the texts. This group comprises Groups (4), (5) and (6). The lexical items in these three groups match what has been discussed in literary criticism of Jane Austen's novels and writing style. For instance, that her novels deal with women's lives in Regency England can be represented by the keyness of words related to women (Group 4). Group (5), which consists of the keyword "sister" and the semantic field "kin", corresponds to the point, noted by Sherry (1966) and Page (1972), that Jane Austen's novels are primarily domestic. Group (6), which contains the keyword "feelings" and various semantic fields (see above), reflects the point critics often make that her novels feature characters' thoughts and feelings, rather than physical actions or adventures. The fact that these corpus-informed sets of findings can be interpreted in close relation to what Jane Austen literary scholars and critics have

been talking about suggests that the corpus-driven keyword analysis conducted in this study can provide textual evidence for previous studies on Jane Austen. Because these three groups of linguistic markers of Jane Austen's writing style have often been mentioned in literary studies of Jane Austen, the present study will not deal with them in detail.

The other group consists of lexical items, as Scott and Tribble (2006) state, that are not usually identified consciously by readers as important but occur in significantly high frequencies and so can be indicators of the style of a text, rather than of its content. Given that Jane Austen's novels are not about the degree of something, we can say that words in Group (1) "words related to a high degree" are style markers, rather than "aboutness" indicators while the other two groups, "modal auxiliary verbs" and "auxiliary BE and HAVE" are helping verbs which generally do not express the content of the texts. To the best of my knowledge, the auxiliary verbs BE and HAVE have never been mentioned anywhere, even in passing, in literary studies of Jane Austen. As for words related to a high degree, they have been mentioned in passing in relation to Jane Austen's characterization of comic and insensible characters through their idiolects (cf. e.g. Booth 1991 and Stokes 1991), i.e. those characters tend to use such intensifiers as "extremely" and "vastly" when expressing their opinions about something. Modal auxiliary verbs have been studied in great detail by Burrows (1986), but they are treated as properties of the speech of characters with a strong sense of morality, e.g. Fanny Price in *Mansfield Park* or Mr. Knightley in *Emma*, while it is revealed in the present study that they are not just characteristic of some characters' speech but style markers of all Jane Austen's six major novels. Since these three groups of findings have not yet received much attention in literary discussions of Jane Austen's works, they will be investigated in turn in the present study.

(1) Words related to a high degree

Of all the six groups of key linguistic features shown above, words related to a high degree are the most characteristic of Jane Austen's novels. This is reflected in the fact that they occur in all three different linguistic categories, as shown below:

Key semantic fields	*degree boosters* *entire; maximum* *degree maximizers*
Keywords	*very* *every* *such* *any*
Key grammatical categories	*degree adverb*

Not only the density of words related to a high degree but also the degree of their keyness reflect their greater significance to Jane Austen's novels than other key features; the semantic field "degree boosters" and grammatical category "degree adverbs" are ranked first in the relevant lists (see Tables 2 and 3 above) while the keyword "very" is in second place in the keyword list (see Table 1 above).

Upon examination of concordance lines of the words in this group, it is found that the words denoting a high degree are used in close proximity to one another. A strong density of high-degree words at some

points in the novels constitutes an exaggerated discourse in Jane Austen's works. The exaggeration, in turn, is likely to encourage readers to feel that the part of the text they are reading cannot be interpreted at face value. Some doubt is likely cast on the reliability or sincerity of the character whose point of view is focalized. The excerpt below, taken from *Emma*, illustrates this point. In this part of the novel, Emma, the protagonist of the novel, has met with Harriet Smith, an orphan from the lower class, for the first time and is interested in making Harriet her protégée. The extract below shows what she thinks about Harriet. The underlined words are those found either in the above-mentioned semantic fields, grammatical categories or keyword list.

> *Emma was as much pleased with her [Harriet Smith] manners as her person, and quite determined to continue the acquaintance. She was not struck by anything remarkably clever in Miss Smith's conversation, but she found her altogether very engaging -- not inconveniently shy, not unwilling to talk – and yet so far from pushing, showing so proper and becoming a deference, seeming so pleasantly grateful for being admitted to Hartfield [Emma's mansion], and so artlessly impressed by the appearance of everything in so superior a style to what she had been used to, that she must have good sense and deserve encouragement.*

In the above paragraph, Emma's satisfaction with Harriet on their first meeting is presented as remarkably strong, as can be seen from the recurrence of high-degree words, especially that of "so". On the surface, Emma may seem very kind and compassionate but her intense admiration for Harriet's manners that show "deference" and "grateful[ness]" for Emma is likely to raise doubt in some readers' minds as to whether Emma's kindness for Harriet comes from her genuine good wishes for Harriet or from her own vanity. In fact, some critics even argue that Emma's decision to adopt Harriet as her protégée is not out of compassion but due to her preference for dominating someone and exercising power (cf. e.g. Mudrick 1952). Of course, the above paragraph does not state so and yet the fact that there are such interpretative arguments and that we find it hard to follow Emma's viewpoint that Harriet is super-good suggests that there are meanings between the lines here and what partly accounts for this textually is the reiteration of words denoting a high degree that present Emma's extreme evaluation.

It must be noted that my attention to the above extract came before I discovered what literary scholars and critics have said about Emma and Harriet. The text was chosen for careful study because concordance lines of words in the above semantic fields show that this part of the novel is rich in hyperbolic words. It can thus be said that concordance investigation is useful in helping stylistic analysts select a text for further detailed qualitative analysis with less subjectivity.[9] This is because the analyst's selection of an excerpt is not guided alone by his/her knowledge and interpretation of a relevant part of the literary work but also through quantitative corpus-driven findings,

[9] Stylisticians are sometimes criticized for being subjective, selecting a text that they already know contains some interesting linguistic features relatable to some interpretative issues (cf. e.g. Fish 1996)

without which he/she may not have at all considered that part of the work.

Apart from the use of high-degree words in the narrative part, it is also found from the concordance lines analysis that this group of lexical items occurs significantly as part of the conversations among characters. It is found from the analysis that, like the narrative part, a character's direct speech that displays a density of high-degree words is a marked exaggeration. The exaggerated speech tends to betray the speaker's insincerity or insensibility. Below is a direct speech quotation of Lucy Steele, an antagonist in *Sense and Sensibility*. In this extract, Lucy is talking to Elinor Dashwood, trying to convince Elinor that she truly loves Edward and it is not because of his prospect of inheriting a large fortune from his mother when she dies.

> *He [Edward] has only two thousand pounds of his own; it would be madness to marry upon that, though for my own part, I could give up every prospect of more without a sigh. I have been always used to a very small income, and could struggle with any poverty for him.*

Like the passage from *Emma*, this extract does not state that Lucy Steele is lying but readers are encouraged not to believe what she says, more or less because her love for Edward sounds exaggerated and that is evidenced by the use of such high-degree lexical items as "very", "every" and "always" close to one another.

To summarise, based on the corpus-driven approach, words denoting a high-degree are found to be most characteristic of Jane Austen's novels. Their close distribution in all her major novels plays a crucial role in suggesting that there are meanings between the lines in many parts of the novels. The word "crucial" used here is not an exaggeration, however, given that the creation and interpretation of meanings between the lines is one of the remarkable qualities of Jane Austen's novels. Literary critics often note that it is not only the female protagonists in all her novels but also her readers who are involved in the process of distinguishing between appearance and reality (McMaster 1996). Decoding meaning between the lines, be it irony or the insincerity of some characters, is a task that her readers will experience in the course of their reading. While this is widely remarked on in studies of Jane Austen, it is hardly ever mentioned in what ways this thematic instantiation is achieved textually. The corpus-driven approach has directed our attention to high-degree words and led us to see that it is this group of lexical items that are used strategically for creating and hinting at meanings between the lines in her novels.

(2) Modal auxiliary verbs

Modal auxiliary verbs are also highly significant to Jane Austen's writing style, whether as a semantic and grammatical group or as a single lexical item, since they occur in all three linguistic levels:

Key semantic fields	*likely* *strong obligation or necessity*
Keywords	*could* *Must*
Key grammatical categories	*modal auxiliary*

The modal verbs that are most characteristic of Jane Austen's novels are "could" and "must"; however, the fact that

modal auxiliary verbs also turn up as key grammatical categories suggests that other modal verbs are also used significantly in her works but they do not turn up above the cut-off point set at LL 200. As for the key semantic fields, although not all the words in the "likely" and "strong obligation or necessity" groups are modal verbs[10], it can be argued that the keyness of these two semantic fields is largely constituted by modal auxiliary verbs since 78.25% of the "likely" semantic field is made up of "could", "would", "can", "might" and "may" while 66.57% of the "strong obligation or necessity" field is made up of "must" and "should".

The textual functions of the modal auxiliary verbs in Jane Austen's novels cannot be approached in the same way as the hyperbolic words analysed above. This is because meanings of the modal verbs vary across the contexts of their occurrences while uses of those high-degree words are closely similar to one another. Therefore, an account of the semantic and pragmatic significance of all modal verbs in Jane Austen's novels cannot be presented here. However, two modal verbs, "could" and "must", will be discussed in some detail because, among all the modal auxiliary verbs, they are also listed as keywords in Jane Austen's novels.

2.1 "could"

The modal verb "could" is found to occur 3,599 times in all six novels. To do a qualitative analysis of textual functions of such a frequent word, reading through all the 3,599 concordance lines of the verb would hardly be possible. Therefore, an automatic extraction of the most frequent phraseological patterns of the verb "could" is opted for, so that it is possible to see in what sort of textual environment the verb is predominantly used. Based on the extraction of three-word clusters of which "could" is a part, which occur more than 50 times in Jane Austen's novels, a total of nine clusters turn up in the list presented below, with the frequency of each cluster in the parentheses.

Top 9 three-word clusters of "could"

1. she could not (333)
2. could not be (167)
3. I could not (87)
4. could not have (76)
5. could not help (76)
6. that she could (75)
7. could not but (73)
8. he could not (69)
9. as she could (66)

As can be seen, out of the nine 3-word clusters, seven of them contain the negative word "not". We can thus infer that the keyword "could" is used repeatedly in the novel to convey characters' inability to do something. Upon further investigation of the concordance lines of all the "could not" clusters, it is found that they tend to co-occur with words or phrases related to cognition, perception and speech, as illustrated below:

- She could not avoid a little suspicion at the total suspension of Isabella's impatient desire to see Mr. Tilney. (*Northanger Abbey*)

- With all these circumstances, recollections and feelings, she could not hear that

[10] The "likely" group also contains such words as "probably", "promising" and "probable" and the "strong obligation or necessity" group consists of such words as "necessary", "obligation" and "duties".

Captain Wentworth's sister was likely to live at Kellynch without a revival of former pain. (*Persuasion*)

- Whether he had felt more of pain or of pleasure in seeing her she could not tell, but he certainly had not seen her with composure. (*Pride and Prejudice*)

This collocational pattern points to one of the central features of Jane Austen's novels, noted by literary critics (e.g. Stovel and Gregg 2002), that in Jane Austen's fictional world, her characters are primarily engaged in conversation and judging others from each character's talk. The density of the "could not" cluster in collocation with verbs of perception, cognition and speech reflects the difficulty the characters, mostly the female protagonists, have in understanding or expressing their thoughts about certain matters. The inability to see through things or speak up is the main problem Jane Austen's protagonists encounter and must be able to solve before they have a happy ending at the close of the novel: Catherine Morland in *Northanger Abbey*, Marianne Dashwood in *Sense and Sensibility*, Elizabeth Bennet in *Pride and Prejudice* and Emma Woodhouse in *Emma* have to learn to perceive the discrepancy between appearance and reality while Elinor Dashwood in *Sense and Sensibility*, Fanny Price in *Mansfield Park* and Anne Elliot in *Persuasion* can see through things and judge correctly but cannot speak their mind because of the force of certain circumstances. The statistically significant predominance of the modal verb "could" can thus be related to the author's portrayal of the problems each character suffers, as shown by the frequent collocation of "could" with "not" and cognition/ perception or speech verbs.

2.2 "must"

The other modal verb that ranks among keywords in Jane Austen's novels is "must", with 2,079 tokens in JA. Like "could", it is hardly possible to investigate all the concordance lines of "must", given its frequency. Therefore, three-word clusters in which "must" is embedded were extracted in order to see the ways in which "must" is often used in Jane Austen's novels. However, the criteria set for extraction of "could" clusters, i.e. the clusters with a minimum frequency of 50, cannot be applied to that of "must", since there are only two clusters that occur more than 50 times (see below). Hence, the ten most frequent clusters of "must" are considered so that more patterns can be found and explored in detail. Below is a list of the top 10 three-word clusters of "must". Note that there are two clusters that have the same frequency. Like "could" clusters, the frequency of each cluster is in parentheses.

Top 10 three-word clusters of "must"

1. it must be (78)
2. must have been (70)
3. must be a (43)
4. must not be (42)
5. he must be (27)
6. you must be (25)
7. you must have (25)
8. must be the (24)
9. that she must (22)
10. I must have (20)

Given the above list, it can be seen that "must" tends to co-occur with two auxiliary verbs, the infinitives "be" and "have" and the past participle "been", with either "be" or "been" embedded in the first seven clusters of "must". Upon investigation of these clusters, it is found that they are often used in the characters'

speculations about certain people or states of affairs.[11] This can be illustrated by the sample concordance lines below, where the relevant clusters are underlined:

- The fact [that Colonel Brandon fell in love with Marianne Dashwood] was ascertained by his listening to her again. It must be so. She was perfectly convinced of it. (*Sense and Sensibility*)

- "No doubt she [Miss Crawford] will be very glad. It must be a great relief to her," said Fanny, trying for greater warmth of manner. (*Mansfield Park*)

- She felt that something must be the matter. The change was indubitable. The difference between his present air and what had been in the Octagon Room was strikingly great. (*Persuasion*)

This pattern of the modal "must" suggests that Jane Austen's novels are to a large extent concerned with characters' speculations about others or certain states of affairs. This relates to the point discussed above in the analysis of "could" clusters: just as the frequent negative clusters of "could" contribute to the description of the characters' inability to understand some people or matters, the clusters of "must" indicate that, in Jane Austen's fictional world, the characters often do not have a clear idea about other characters or certain matters and hence have to guess what exactly is the case. These two keywords therefore serve as linguistic evidence that accounts for literary critics' interpretation that Jane Austen's works tend to be lacking in physical action but are full of the narrator's portrayal of the characters' thoughts and presentation of their conversations.

(3) BE and HAVE

Table 1 above shows that two verb forms of the lemma BE, "be" and "been", and the infinitive "have" are among the top 17 keywords and key grammatical features in Jane Austen's novels. However, they do not occur in the key semantic domains. This is probably because, based on an investigation of the concordance lines of the verbs, many of them are used as helping verbs, whose role has more to do with grammatical form than semantic or pragmatic aspects of the texts. Although they are mainly used as helping verbs, the statistical significance of their occurrences in the novels should lead us to turn our eyes to them and find out why such small words mark the writing style of this great author.

Since the three verb forms occur considerably in JA[12], it is hardly possible to analyse every single concordance line of the verbs. It is therefore more helpful to extract predominant patterns in which BE and HAVE occur. To this end, an extraction of statistically significant

[11] Another use of "must", though less frequent, is suggested by the cluster "must not be" and "that she must", which rank fourth and eighth, respectively, in the list. Unlike the other clusters, "must not be" tends to be used in a character's speech when the speaker, often higher in status or older than the interlocutor, indicates that he/she does not want something to happen. For example:

- "My dear friend, you must not be angry with me" (*Northanger Abbey*)

- "You must not be too severe upon yourself" (*Pride and Prejudice*)

[12] In the six novels by Jane Austen, there are 8,157 tokens of "be", 3,257 tokens of "been" and 5,189 tokens of "have".

collocates of the two verbs was conducted. It should be noted that I did not choose to extract frequent clusters as in the analysis of "could" and "must" because, based on the experiments with the Cluster function on WordSmith, the frequent clusters of the two auxiliary verbs tend to display the co-occurrence among auxiliary verbs, such as "will have been", which requires still further steps in the analysis. An extraction of collocates of "be", "been" and "have", which, unlike cluster analysis, includes words that do not necessarily occur immediately before or after the verb forms but still occur within the four-word span of the node words, seems to be more helpful in showing phraseological patterns of the three verb forms. To extract significant collocates of "be", "been" and "have", three criteria were set up as follows:

(1) The collocates must be lexical items with a minimum frequency of 30 tokens in JA

(2) The collocates must be lexical items found to occur within the 4-word span to the right and left of the search word

(3) The collocates must have a minimum statistical MI value of 3.

The collocates of these three keywords can similarly be divided into seven groups: (1) modal auxiliary verbs, (2) personal pronouns, (3) prepositions (i.e. "by" and "before", (4) adjectives, (5) adverbs, (6) lexical verbs and (7) nouns.

However, the dominant group of collocates of each verb varies. The largest group of collocates of "be" is adjectives, namely "glad", "sorry", "satisfied", "sure", "happy", "able", "likely", and "better".[13] As can be seen, many of these adjective collocates of "be" are concerned with thoughts and feelings. This suggests that the auxiliary verb "be" and its collocation with adjectives denoting thoughts and feelings is a statistically significant collocational pattern in Jane Austen's novels. This reflects that what features in the author's novels is the description of characters' thoughts and feelings, rather than their physical actions. In fact, as shown in Table 3 above, lexical items about thoughts and feelings constitute the key semantic domains in Jane Austen's novels. Although the other three adjectives, "able", "likely" and "better", are not directly about thoughts and feelings, an investigation of their concordance lines suggests that they are also more or less connected to thinking and feeling since they are part of the characters' judgment or evaluation of others or certain states of affairs. This is illustrated in the following sample concordance lines of the collocation among "be", "likely" or "able" or "better", and evaluative words or phrases:

- Their drive, even when this subject was over, was not likely to be very agreeable. (*Northanger Abbey*)

- But I shall tell you, Miss Anne, because you may be able to set things to rights, that I have no very good opinion of Mrs. Charles' nursery maid. (*Persuasion*)

- She […] thought it would be better to speak openly to her aunt than to run such a risk. (*Pride and Prejudice*)

[13] There are relatively much fewer cases in which "better" is used as an adverb, compared with its use as an adjective.

Taking all the above statistically significant adjective collocates together, we can see that the pattern of "be + adjective" is the dominant phraseological pattern that is connected to the noted quality of Jane Austen's writing style, i.e. her works mainly involve characters' judging someone or something.

Unlike "be", the most predominant group of collocates of "been", the other verb form of BE ranked among the keyword list, is adverbs, namely "never", "always", "ever", "too", "so" and "much". All of these collocates can be used to express a high degree of something. This collocational pattern of "been" and high-degree adverbs is linked to the creation of exaggerated discourse in Jane Austen's works. As already discussed above, these hyperbolic statements perform crucial textual functions in the novels. For example, they may be a part of a character' speech, which often betrays the speaker's extreme, unreliable judgment or insincerity, or articulates an ironic force in the narrative presentation of a character's thoughts, or the narrative description of some character or event. This is illustrated through the following sample concordance lines:

[Context: Lucy Steele is trying to mislead Elinor that Edward loves her very much.]

- I have never been able," continued Lucy, "to give him my picture in return, which I am very much vexed at, for he has been always so anxious to get it! (*Sense and Sensibility*)

[Context: Emma is pondering that after she and Frank Churchill have not seen each other for a while, she has lost her feelings for him and he would be upset if he found this out. But, in fact, Emma does not know at all that Frank has always only pretended to feel attached to her.]

- Her own attachment had really subsided into a mere nothing; it was not worth thinking of; -- but if he, who had undoubtedly been always so much the most in love of the two, were to be returning with the same warmth of the sentiment which he had taken away, it would be very distressing. (*Emma*)

[Context: While other characters go out, Mrs. Rushworth and Mrs. Norris are left at home. However, this is not a problem for Mrs. Norris as she likes to flatter the rich, such as Mrs. Rushworth. In this excerpt, the narrator describes her enjoyment sarcastically.]

- Mrs. Norris had been too well employed to move faster. Whatever cross-accidents had occurred to intercept the pleasure of her nieces, she had found a morning of complete enjoyment (*Mansfield Park*)

As for the verb "have", its largest group of collocates is lexical verbs in the past participle form, namely "seen", "heard", "known", "thought", "given", "done", and "made". This does not appear very surprising, given that the node word is the verb "have" and hence the collocational pattern between "have" and past participle verbs reflects the grammatical construction of the present perfect. However, upon investigation of the list of those significant verb collocates, it is observed that they have in common certain semantic properties; that is, the verbs "seen" and "heard" are verbs of perception and "known" and "thought" are verbs of cognition. The other three verbs, "given" and "made", though not conveying meanings of perception or cognition, are in

many cases used with words related to thoughts and feelings. This is illustrated below with the relevant words underlined:

- I think differently now; time and sickness and sorrow have given me other notions. (*Persuasion*)

- Could she but have given Harriet her feelings about it all? She has talked her into love; but, alas! she was not so easily to be talked out of it. (*Emma*)

- The letters from town, which a few days before would have made every nerve in Elinor's body thrill with transport, now arrived […] (*Sense and Sensibility*)

- But your arts and allurements may, in a moment of infatuation, have made him forget what he owes to himself and to all his family. (*Pride and Prejudice*)

- But with sense and temper which ought to have made him judge and feel better, he allowed himself great latitude on such points. (*Mansfield Park*)

- "And now, Henry," said Miss Tilney, "that you have made us understand each other, you may as well make Miss Morland understand yourself […]" (*Northanger Abbey*)

The fact that the keyword auxiliary verb "have" tends to be used in collocation with words related to perception, thoughts and feelings, serves as a set of linguistic evidence that accounts for the reason why literary critics tend to feel that Jane Austen's novels are lacking in action.

The collocational pattern of "have" and "done", however, occurs significantly in the six novels because they are often used as part of the narrative description of conflicts between characters' words or actions and their thoughts or contrasts between what a character speculates and what really happens. This is reflected in the fact that "have done" tends to co-occur with modal verbs, as shown in the sample concordance lines below:

- "Well, Catherine, how do you like my friend Thorpe?" Instead of answering, as she probably would have done, had there been no friendship and no flattery in the case, "I do not like him at all," she directly replied, "I like him very much; he seems very agreeable." (*Northanger Abbey*)

- Had he wished ever to see her again, he need not have waited till this time; he would have done what she could not but believe that in his place she should have done long ago, […] (*Persuasion*)

- "[…] However, I recollected afterwards that if he had been prevented from going, the wedding need not be put off, for Mr. Darcy might have done as well." (*Pride and Prejudice*)

Given such repeated uses of the phraseological pattern "modal verb + have + done" in Jane Austen's novels, it can be said that the recurrence of this pattern contributes to the interpretation that Jane Austen's novels deal with differences between appearance and reality.

Discussion

The findings presented above throw light on two important points in relation to the research questions (see above) addressed in the present study: (1) textual patterns and their relationship to meaning in Jane Austen's novels and (2) assessment of a corpus-driven approach to the study of

literary texts. These two points will be discussed together in this section.

By comparing Jane Austen's novels with a corpus of their late 18th – early 19th century contemporaries and with a corpus of present-day British fictional prose, it is found that the lists of statistically significant lexical items, grammatical categories and semantic fields in Jane Austen's novels correlate with one another to a large extent. This suggests that, whether we take a lexical, grammatical or semantic perspective, six groups of linguistic features are particularly characteristic of Jane Austen's novels, namely:

(1) words related to a high degree,

(2) modal auxiliary verbs,

(3) the auxiliary verbs BE and HAVE,

(4) words related to internal states of mind,

(5) words related to family relationships,

(6) words related to women.

In the light of what has been discussed in literary studies, some of the above corpus-driven findings, especially groups (4) – (6), cannot be said to be totally new or to throw fresh light on stylistic features of Jane Austen's novels, since they have been already mentioned, more or less, by literary critics. In fact, to the best of my knowledge, only group (3) "the auxiliary verbs BE and HAVE" has not been talked of in any previous studies of Jane Austen. On the surface, this might be taken to mean that a corpus approach does not seem very helpful in the study of literature, in this case Jane Austen's novels, since important textual features can be observed through scholars' close readings. However, we should not forget that those claims are intuition-based and that they can now be validated (or refuted) through the findings derived from a quantitative comparative method. In short, though not providing a totally new set of findings or generating new points of discussion, a corpus approach can provide statistically significant textual evidence that helps support or refute claims made in literary studies.

Having said that, I must admit that I have some reservations about the value of some textual evidence derived from the corpus-driven method. Looking, for example, at the density of words related to "women" and "family relations", I cannot help but wonder whether we need a corpus approach to explain that Jane Austen's novels deal with women and their families. Is it a worthwhile effort to compile a corpus, reference corpora, and conduct a statistical quantification, just to find that Jane Austen's novels are primarily concerned with women and family matters? I personally feel that only some sets of corpus findings, to be discussed below, can be of value to literary criticism while others, such as the keyness of words related to women, seem to point to aspects that are too general or superficial for literary discussion. This may be because, unlike academic or other informative texts, literary texts are expressive texts, whose thematic meanings are not always conveyed straightforwardly through the words that appear on the surface of the text. The keywords that are content words, which generally indicate the "aboutness" of a text as Scott and Tribble (2006) state, do not seem to be of much value to a stylistic study of literary texts.

A more valuable set of findings that a corpus-informed method yields, in my opinion, is those that involve function words or semi-grammatical lexical items, such as lexical items in groups (1) – (3). This is because this group of findings can hardly ever be detected even by the most careful readers. To me, the statistical significance of the auxiliary verbs "be", "been" and "have" and their collocational patterns in the texts is a prime example that illustrates the value of a corpus-driven method in shedding light on linguistic features and patterns that influence our interpretations but are very likely to escape scholarly attention.

This also applies to the linguistic categories that have been less observed by literary critics. While the high-degree words are rarely mentioned in literary criticism of Jane Austen (and when they are, they often occur in passing), it is the corpus-driven approach in this study that enables us to find out that these apparently small words turn out to be the most distinctive stylistic feature of Jane Austen's writing. Moreover, while what has been mentioned by critics are all marked intensifiers, such as "vastly" and "perfectly", it has been revealed through a corpus-driven approach in this study that it is not simply the use of intensifiers but also other sorts of words denoting a high degree, e.g. adverbs of frequency like "always" and "never" and indefinite pronouns like "everybody", "anyone", that are used significantly and, more importantly, in close proximity to one another. In other words, the corpus approach has shown that it is not just the use of intensifiers but also the close proximity of high-degree words of various kinds that serves as a tool for the author to create and hint at meanings between the lines in her novels.

In the case of modal auxiliary verbs, though having been studied by Burrows (1986), the corpus-driven approach has enabled us to explicate their roles in Jane Austen's novels in a more precise and refined manner. It shows patterns of co-occurrences between "could" and "not" or "must" and "be", which in turn helps illuminate the instantiation of thematic ideas in her novels.

Based on my observation of the set of findings from the study, it can be said that a corpus-driven approach is of value to a stylistic analysis of literary texts at varying levels. At the most basic level, it can be a "supporting actor" in literary or stylistic research, yielding quantitative textual evidence that supports or refutes intuitive interpretations. More than that, a corpus approach can be the "main actor" that unearths linguistic or stylistic features that even a well-trained reader could hardly imagine in explaining interpretative issues. Finally, while it has been widely acknowledged that a corpus approach seems to be the only method analysts can resort to if a whole work of fictional prose or a group of literary works are objects of the study, it has been found from the present study that the corpus-driven approach can also be of great help for a detailed manual analysis of part of the whole text, since it can draw our attention to part(s) of a literary work that is(are) worth further detailed investigation, as illustrated in the analysis of high-degree words.

Despite such potential, the comparative corpus-driven technique has certain limitations. First, since corpus linguistics holds that the more frequent linguistic patterns are, the more significant they are, an application of a corpus-driven approach in stylistics relies heavily on quantitative

value, when, in fact, items or patterns that lie above the cut-off point may not be of much importance in a literary text. For example, as discussed above, the keyness of words related to women and family ranks very high and yet these words are not very illuminating textual features when it comes to an academic discussion of Jane Austen's works. On the other hand, items that are below a statistical cut-off point can be significant to an analysis of a literary text. In fact, as Leech and Short (1981) argue, a word that occurs only once in a text may be of great significance to the text under investigation. Second, the corpus approach allows us to explore mainly lexical items and their patterns in a text. Other linguistic aspects, such as characters' interactions, can hardly be dealt with from a corpus-stylistic approach. Irony in Jane Austen's novels is a good example that illustrates this. Although the present study enables us to see that an extensive distribution of various kinds of hyperbolic words plays an important role in creating and interpreting meanings between the lines in Jane Austen's novels, it is undeniable that the understanding of irony requires more than the recognition of high-degree words. A pragmatic or cognitive perspective must also be involved in a full explanation of ironic statements. In other words, the corpus-driven approach offers an insight into only one aspect of textual features. Finally, a word of caution is in order. The corpus-driven approach simply shows us ***what*** is statistically significant in the text, it does not help explain ***to what extent*** and ***why*** the lexical item is significant; that job is the analyst's.

Conclusion

The present study starts from an observation that while a corpus linguistic technique has been adopted in a number of types and forms of discourse and text analyses, it has been adopted only infrequently in a stylistic analysis of literary texts. A corpus-driven approach was thereby applied to an analysis of Jane Austen's six major novels in order to see how well this method works with literary texts. Although I did not identify in the first place what linguistic features should mark the style of Jane Austen's works, the corpus-driven method has yielded a set of findings that are useful for the discussion of Jane Austen. Some can help support literary scholars' observations on Austen's novels in quantitative terms while others serve as new linguistic evidence that can enrich the study of the novelist. However, although reliance on a computer makes it possible to investigate a group of literary texts and provides satisfactory results, it is only part of the story; the analyst's understanding of the text in question still plays a central role in explaining and evaluating those corpus-informed findings.

References

Austen, Jane. *Emma*. London: Penguin Books, 1994.

---. *Mansfield Park*. Harmondsworth, Middlesex: Penguin Books, 1985.

---. *Northanger Abbey*. New York: Bantam Books, 1989.

---. *Persuasion*. New York: Bantan Books, 1989.

---. *Pride and Prejudice*. London: Penguin Books, 1994.

---. *Sense and Sensibility*. London: Penguin Books, 1994.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London/ New York: Continuum.

Biber, Douglas and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In *Out of corpora*, edited by H. Hasselgard, & S. Oksefjell, pp.181-190. Amsterdam-Atlanta GA: Rodopi.

Biber, Douglas. 2011. Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature* 1(1), 15-23.

Booth, Wayne. 1991. Control of Distance in Jane Austen's *Emma*. In *Jane Austen: Emma (A Casebook)*, edited by D. Lodge. London: Macmillan Press, 137-55.

Burrows, J. F. 1986. Modal verbs and moral principles: An aspect of Jane Austen's style. *Literary and Linguistic Computing* 1, 9–23.

Church, Kenneth and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16 (1), 22-29.

Enkvist, Nils. 1973. *Linguistic Stylistics*. The Hague: Mouton.

Fish, Stanley. 1996. What is Stylistics and Why are They Saying Such Terrible Things About It? In *The Stylistics Reader*, edited by J. Weber, 94-116. London*:* Arnold.

Fletcher, William. 2003. April 12, 2011 <http://phrasesinenglish.org/>

Hoey, Michael. 1991. *Patterns of Lexis in Text*. Oxford: Oxford UP.

Leech, G. and Short, M. 1981. *Style in Fiction*. London: Longman.

Leech, G., Rayson, P. and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English Based on the British National Corpus*. London: Longman.

Mahlberg, Michaela. 2005. *English General Nouns: A Corpus Theoretical Approach*. Amsterdam: John Benjamins.

--- 2007. Clusters, key clusters and local textual functions in Dickens. *Corpora*, 2(1), 1-31.

McMaster, Juliet. 1996. *Jane Austen the Novelist: Essays Past and Present*. Basingstoke: Macmillan.

Moon, R. and Carmen. R Caldas-Coulthard. 2010. 'Curvy, hunky, kinky': using corpora as tools for critical analysis. *Discourse & Society*, 21(2) 99–133.

Mudrick, M. 1952. *Jane Austen: Irony as Defense and Discovery*. Princeton: Princeton UP.

Page, Norman. 1972. *The Language of Jane Austen*. Oxford: Basil Blackwell.

Rayson, Paul. 2007. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University, January 23, 2011. <http://www.comp.lancs.ac.uk/ucrel/wmatrix/>

Scott, Mike. 1999. WordSmith Tools (Software). Oxford: Oxford UP.

Scott, M. and Christopher Tribble. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language Education*. The Hague: John Benjamins.

Sherry, Norman. 1966. *Jane Austen*. London: Evans Bros.

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

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Stokes, Myra. 1991. *The Language of Jane Austen: A Study of Some Aspects of Her Vocabulary*. Basingstoke: Macmillan.

Stovel, B. and Lynn Gregg. 2002. *The Talk in Jane Austen*. Alberta: U. of Alberta Press.

Stubbs. M. 2005. Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*, 14, 1: 5-24.

Stubbs, M. and Isabelle Barth. 2003. Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language*. 10, 1: 65-108.

Tanner, Tony. 2007. *Jane Austen* (reissued edition). Hampshire: Palgrave Macmillan.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.