DALARNA UNIVERSITY

Degree Project

Level: bachelor's

The Female Protagonists in Thackeray's Vanity Fair

A Corpus Linguistic Study of Keywords, Collocations, and Characterisation

Author: Tina Åhman Billing
Supervisor: Julie Skogs
Examiner: Jonathan White
Subject/main field of study: English Linguistics
Course code: EN 2035
Credits: 15
Date of examination:


DALARNA
UNIVERSITY

Abstract

This essay uses corpus linguistic methods to study aspects of the novel *Vanity Fair* by W M Thackeray. The aim is to study the way Thackeray chose to describe his two female protagonists, Rebecca Sharp and Amelia Sedley. This is accomplished by a closer study of keywords in *Vanity Fair*, created by using a reference corpus consisting of thirteen novels by Victorian authors. These keywords are used to define semantic fields related to the novel. Keywords from the semantic field closest to the protagonists are studied in context. In addition, adjectives that collocate with the names of the protagonists are analyzed to compare the characterization of each woman. The study indicates that Thackeray has used fewer adjectives to describe Amelia than Rebecca, but that he has used these more frequently, which may cause readers to form a stronger mental picture of Amelia's character sooner than they do for Rebecca's.

*Keywords***:** Corpus Linguistics, Corpus Stylistics, Characterisation, Thackeray, Vanity Fair

Table of contents

1 Introduction

With the advent of affordable computer processing and storage, Corpus Linguistics has come a long way from the manual handling of (a limited number of) paper slips in shoeboxes to the massive collections of texts that can be processed electronically in the blink of an eye. This development has allowed corpus based methodologies to be used more broadly than just for "pure" grammar or lexical studies (McEnery, Xiao, & Tono, 2006, p.3). In the last decade or so, a new research area, corpus stylistics, has developed out of corpus linguistics. Corpus stylistics focuses on literary analysis supported by corpus linguistic methods.

There are many quite recent examples of corpus stylistic studies of the works of e g Austen, Shakespeare, Dickens, Woolf, and Salinger to name a few (Fischer-Starcke 2010, Leech 2011, Mahlberg 2013, Balossi 2014, Yazdanjoo, Sabbagh, & Shahriari 2016). However, there seem to be no published corpus stylistic studies on the works of W M Thackeray. His novel *Vanity Fair* has been the subject of literary analysis and critique since its publishing in 1848. Scholars have focused on a variety of themes such as the structure of the narrative, the narrator's position, Thackeray's irony, his possible subversive criticism of Victorian society, or even very specific details such as how one of the protagonists uses her hands (Cecil 1934, Gilmour 1982, Jadwin 1992, Pietka 2010, Cammack 2015).

The subtitle of *Vanity Fair* is *A Novel Without a Hero.* Critics have been discussing whether instead the novel has a heroine. Readers remember Rebecca Sharp as the main character of the story, but how does Thackeray portray her? Do the choices made by the author suggest that he wants her to be seen as a heroine, or as a monster? Litvak is not the only critic who has found it difficult to pinpoint Rebecca's exact status and Thackeray's seemingly ambivalent attitude towards her: "the mixture of admiration and condescension with which he treats her" (Litvak 1996, p. 232).

Rebecca's friend from school, Amelia Sedley, is the second female protagonist in the novel, and the reader's perception of her characteristics may seem to be the inverse to those of Rebecca. "The characters of the two girls are designed to illustrate the laws controlling Vanity Fair as forcibly as possible. And in order to reveal how universally these laws work, they are strongly contrasted types" (Cecil, in Thackeray 1994, p. 815).

This study will use corpus linguistic and corpus stylistic methodology to study some linguistic aspects of *Vanity Fair*. More specifically it will investigate the way in which Thackeray makes use of linguistic features to depict the two main female protagonists in the novel.

1.1 Aim of the Study

The purpose of this study is to compare the way Thackeray chooses to describe the two main female characters in his novel *Vanity Fair*, using corpus linguistic methods, and especially in relation to the main themes of the novel. By identifying words which are used more frequently in *Vanity Fair* in comparison to a reference corpus consisting of other Victorian novels, the main themes specific to *Vanity Fair* can be identified. The comparison will be made using corpus linguistic tools. In addition, some of the words or phrases most commonly used in connection with the two protagonists, Rebecca and Amelia, will be studied in context in order to analyze Thackeray's portrayal of these two characters. The study will be a combination of quantitative and qualitative analyses.

More specifically, the research questions that will be answered are:

- Which keywords are unusually frequent in *Vanity Fair* in comparison with a corpus of Victorian novels and what themes, or semantic fields, can be identified through these keywords?

- How do the semantic fields relate to the two characters, Rebecca and Amelia?
- What linguistic features does Thackeray use to characterise the two female protagonists in *Vanity Fair*?

2 Theoretical Background

This section will describe how corpus stylistics relates to corpus linguistics. It will then go on to introduce some of the aspects relating to characterisation and some of the theories on this topic. Last, it will give a brief description of *Vanity Fair* and its relation to Victorian society.

2.1 Corpus Linguistics and Corpus Stylistics

Today's large corpora, together with corpus tools, make it possible to identify linguistic patterns across different texts and text types. This differs from literary stylistics where the focus is on the specifics of a single text, or even just a text sample (Mahlberg, 2013, p. 7). Corpus stylistics is a field at the intersection between corpus linguistics and literary stylistics. It is the application of corpus linguistic methods to help answer questions raised in the fields of literary stylistics and literary criticism (Mahlberg, 2010, p. 295). Whereas traditional literary stylistics is limited to studying shorter texts or extracts from longer texts, computer-aided corpus stylistics is necessary for the detailed linguistic study of longer texts (Carter 2011, p. 65) as it "allows for the detection of patterns which are invisible to an analyst's naked eye" (Fischer-Starcke, 2009, p. 494). The use of the frequency-based approach enabled by corpus tools makes it possible to avoid the subjective choices of which features to analyse that traditional stylistic analysis depends on. It is instead possible to find "detailed and neutral insights into the data, which are independent of, for example, previous knowledge of the reception of the work or genre conventions" (Fischer-Starcke, 2010, p. 6).

The use of corpus linguistics to help with the analysis of fictional literature is still a relatively new research field. Both Fischer-Starcke and Mahlberg, two of the pioneers in the field, give examples of how corpus linguistic tools can provide new insights as well as support a type of research that has hitherto been seen as far too time-consuming to undertake. Mahlberg emphasizes the need to combine quantitative findings gleaned from the corpus with qualitative analysis (2010, p. 292), but she also makes clear that the corpus linguistic approach to literary analysis takes its starting point in textual patterns, rather than in the reader's cognitive processes (2010, p. 297). The use of corpus linguistics for literary analysis has garnered criticism from some scholars, but Fischer-Starcke argues for the complementary qualities of literary studies and corpus-based studies, as they will "provide different insights into the same texts" (2010, p. 8).

2.2 Characterisation

Readers depend on linguistic forms to infer an idea of the literary protagonists they encounter in a text (van Peer 1989, p. 9). Although literary characters usually play an important part in a narrative, limited effort has been spent on researching what the process of inferring information about them looks like (Culpeper 1996, p. 335). He suggests synthesizing attribution theories from social psychology with literary and linguistic theories (Culpeper 1996, 2000).

In a recent overview of the theories around characterization (with contributions mainly from literature and media studies), Heidbrink summarizes the current status of this field of research and notes that different theoretical perspectives are starting to merge and also combining psychologically- and anthropologically-based approaches into the work. She notes that the reader (or viewer in the case of film) uses the human being as a primary reference for understanding a character. One of the points she makes is that both the role the character plays

in the plot as well as the attributes it is equipped with are of interest. (Heidbrink, 2010, pp. 99-100). Quite in line with the trends described by Heidbrink, linguists Stockwell and Mahlberg proposes a theory of mind-modelling, which takes its starting point in the assumption that each reader begins the process of imagining the characters he reads about with a kind of template for each character, loosely modelled on the reader's self. This template is then modified by the reader according to the information and inferences coming from the text, "until a sufficiently working model of the other person is rendered in mind" (Stockwell & Mahlberg, 2015, p. 133).

In more practical terms, one of the means to draw readers' attention to information that will support their interpretation of the text (and the character) is foregrounding. Foregrounding can take the form of regularity by the establishment of extra patterns, or irregularity by breaking linguistic norms (Culpepper, 1996, p. 346, Leech, 2011, p.16). Foregrounding can be achieved by repetition, as a way to break linguistic norm through overfrequent use of a linguistic feature (Mahlberg, 2013, p.8). Emmott and Alexander suggest that the repetition of characters' attributes may increase the reader's awareness of them and contribute to the storing of this information, as we build our mental representation of the characters (2011, p. 331). The elements in the text used for inferring information about the characters can be more or less explicit in nature. Among them are: direct descriptions, presentation of speech, representation of thoughts, motives, desires and intentions, reactions from other characters or the narrator (Culpeper 1996, Stockwell & Mahlberg, 2015).

Mahlberg (2013) has studied Dickens's use of repeated phrases, or word clusters, as a way to cumulatively build pictures of his characters. These clusters are analysed in relation to character speech and body language. She also looks specifically at collocations of *as if* as these tend to be pointing to features of the characters. Fischer-Starcke (2009) has used keywords as the starting-point to identify important topics in Austen's *Pride and*

Prejudice. Combining these with a study of frequently used four word phrases and their collocations, she goes on to identify some of the characteristics of female speech in the novel, showing how the female characters are hedging their statements whenever they voice disapproval. This, according to Fischer-Starcke, makes the female characters appear strong and independent to the reader, while still adhering to social norms.

In this study, with its limited scope, the focus will mainly be on direct descriptions of characters, but some of the other elements which could be used to infer information about characters may also prove useful for analysis.

2.3 Thackeray, Victorian Society, Vanity Fair, and Satire

During the reign of Queen Victoria, the industrial revolution in Britain paved the way both for the generation and re-distribution of wealth. This made it possible for individuals to aspire to a higher social rank. As a consequence, issues of social standing became central to the Victorians in general, and for Victorian authors social ambition became a common theme to explore in their writing. In *Vanity Fair*, Thackeray chose to tell the story of Rebecca Sharp, an accomplished young woman of simple origin but with strong social ambitions. Her unscrupulous methods for climbing the rungs of the social ladder are contrasted with the passive ways of Amelia Sedley, a friend of Rebecca's from school, who comes from a well-off home.

The structure of the novel is relatively complex, as, among others, Harden (1967 in Thackeray 1994, pp.710-730) has described. The two female protagonists, Rebecca and Amelia, are often contrasted (one is fair and one is dark; one is active, the other passive), and although their stories run in parallel at times, they are often at opposing ends of some scale (Gilmour 1982, p. 16). "Thackeray creates characters who move in and out of [..] categories" according to Pietka (2010, p. 241). With this in mind, it is interesting to find out whether

Thackeray uses linguistic features to convey this sense of his characters, or whether it is solely limited to plot development (and thus an investigation more suitable for traditional literary analysis).

Vanity Fair is known as a satire and Jadwin (1992) among others points out how Thackeray uses language to convey meanings alternative to those on the surface of the text. There is obviously a need to take into consideration the possibility that words or phrases should not always be taken at face value, when using quantitative methods to analyse a text such as *Vanity Fair.* Louw (1993) looks specifically at the possibility of identifying irony in a text. He suggests that irony in a text can be found e g when the semantic prosody of a collocation seems to go against the norm. To be able to identify the established prosodic norm, a substantial corpus is needed (Louw 1993, p. 164). Louw has not been left unchallenged, but with the limited scope of this study, there will be little room to look into possible markers of irony based on semantic prosody. It will, however, be wise to keep in mind that there may be hidden (ironic) meaning in the words or phrases studied and that this will not show up in raw frequency numbers.

2.4 Definitions

Although core terminology will be defined or explained in the running text on their first appearance, a few key terms are explained here for easier reference.

Keywords: In corpus linguistics the term keyword is used to denote words that are of unusually high frequency in the studied corpus compared to their frequency in a reference corpus (Fischer-Starcke 2009, p. 495, Balossi 2014, p. 46).

Collocates: Collocation is defined differently among linguists. Some claim that for lexical items to be considered collocates, they need to frequently co-occur (McEnery et al. 2006, p. 82-83). In this study the term will be used for co-occurring words indiscriminately of the frequency with which they co-occur.

Concordances: "A set of concordance lines is a sample of a node word together with a sample of its linguistics environments, often defined as a span of words to left and right." (Stubbs, 2001, p. 152)

3 Methodology and Data

This section will begin with an overview of how the study will be carried out and will then continue with a presentation of the two corpora that will form the basis of the study. The chapter ends with a short introduction to the concordance tool that is essential to carrying out the research.

3.1 Methodology

This study will apply corpus linguistic methodology on a literary text, *Vanity Fair*. For this purpose, a primary corpus consisting of the text of the novel will be created. This corpus will be used for the text-internal study of the novel.

A reference corpus will be compiled from other literary texts contemporary with *Vanity Fair*. The reference corpus will make it possible to study *Vanity Fair* in comparison to similar texts. By using a corpus tool, it will be possible to find out which words are relatively more frequently used in *Vanity Fair* than in comparable texts, and thus identify the words that can be seen as specific to the novel, its keywords. As the keywords will carry limited information on their own, they will be grouped according to semantic fields, assuming that

keywords from a dominant semantic field will be important for data analysis (Fischer-Starcke 2009, p. 496). In addition, the keywords will be compared to the collocates of the two female characters, to get an indication of which keywords are appearing most frequently in close context with their names (Mahlberg, 2013, p. 131). Some of these keywords will be analysed with the help of a concordance tool. The concordance tool makes it possible to see them in their linguistic context (Balossi, 2014, p. 47).

The text-internal study will focus on words that are of importance for the characterization of the two female protagonists. These will be examined with the help of the concordance tool, both in terms of their frequencies in the text, their collocations, and their linguistic context. The manner in which they contribute to the characterization of the two female protagonists will be analysed.

3.2 Data

The primary data used in this study consists of two corpora. The primary corpus is made up of one text only, Thackeray's *Vanity Fair*. The reference corpus used is the *Victorian Authors Reference Corpus*, which is compiled of thirteen novels written by authors that were contemporaries of Thackeray. The corpora are hereafter referred to as the *VF Corpus* and the *VictA Corpus* respectively. The contents of the two corpora are presented in Table 1 below.

Vanity Fair (VF) Corpus		Victorian Authors Reference (VictA) Corpus	
		Title	Author
W M Thackeray (1811-1863)	*Vanity Fair (1848)*	*Sybil (1845)*	Benjamin Disraeli (1804-1881)
		Jane Eyre (1847)	Charlotte Brontë (1816-1855)
		Wuthering Heights (1848)	Emily Brontë (1818-1848)
		North and South (1854-55)	Elizabeth Gaskell (1810-1865)
		Guy Livingstone (1857)	G A Lawrence (1827-1876)
		Tom Brown's School Days (1857)	Thomas Hughes (1822-1896)
		Doctor Thorne (1858)	Anthony Trollope (1815-1882)
		Great Expectations (1860-61)	Charles Dickens (1812-1870)
		East Lynne (1861)	Ellen Wood (1814-1887)
		Lady Audley's Secret (1862)	M E Braddon (1835-1915)
		Alice's Adventures in Wonderland (1865)	Lewis Carroll (1832-1898)
		The Moonstone (1868)	Wilkie Collins (1824-1889)
		Middlemarch (1871-1872)	George Eliot (1819-1880)
	311,458 tokens		**1,789,815** tokens

Table 1. An overview of the two corpora, Vanity Fair Corpus *and* Victorian Authors Reference Corpus.

All of the texts in the two corpora have been collected from the *Project Gutenberg* website, where they are freely available, as the copyright has expired for all the material published there. The novels in the *VictA Corpus* were chosen using three criteria. Firstly, the texts needed to be available on *Project Gutenberg*. Secondly, they had to be published around the time that *Vanity Fair* was published. Elliot's *Middlemarch*, which is published 23 years after *Vanity Fair*, represents the largest time gap, but is thematically relatively close to *Vanity Fair*. Thirdly, the authors' year of birth should not be too far from that of Thackeray (1811).

It is of course possible to debate whether the novels included in the reference corpus are the correct or best ones to include. For a larger study, it would be desirable to work with a larger reference corpus, since this would lessen the influence which an individual novel could have on the results (Fischer-Starcke, 2009, p. 499). As this study is limited in scope, the results should only be seen as indicative rather than absolute.

3.3 The Concordance Tool

The concordance tool used for the study is AntConc, version 3.4.4 for Windows, created by Laurence Anthony. The tool is freely available on the internet.

4 Data Analysis and Results

This section will present the data analyses. The outcome of these analyses will be discussed in conjunction with the presentation of some of the data. The section starts with a presentation and discussion about the various name forms used in the novel. It continues with a section focusing on keywords and semantic fields. The third section discusses collocates and attributions.

4.1 Alternative Naming

In order to be able to use the concordance tool for a closer study of the two young women and their positions in the text, their alternative names need to be taken into consideration. They both have a nickname as well as a maiden name and then new names after they are married. In addition, Rebecca's name when married, Mrs. Crawley, is shared with another member of her new family, as is Amelia's (Mrs. Osborne). All the occurrences of these name forms therefore need to be examined in context to ensure that only those referring to the two protagonists are

included in the analysis. Table 2 shows the number of occurrences for each of the name forms after the removal of references to characters other than the two main protagonists.

Name forms - Rebecca	Freq	Name forms - Amelia	Freq
Miss Sharp	165	Miss Sedley	50
Rebecca	511	Amelia	638
Becky	421	Emmy	194
Mrs. Crawley*	95	Mrs. Osborne **	69
Mrs. Rawdon Crawley	30	Mrs. George Osborne	13
Total	**1222**		**964**

Table 2. Frequency of the various name forms for the two female protagonists in Vanity Fair.

**There is more than one Mrs.Crawley in the novel. The mentions referring to Rebecca have been manually identified in context.*

***There is more than one Mrs. Osborne in the novel. One mention referring to George Osborne's mother has been manually identified in context.*

Judging by the number of mentions in the novel, it seems that Rebecca is the more prominent of the two. She is referred to by name a little more than 25% more often than Amelia is. This will not necessarily mean that Amelia occupies less room in the narrative, but it could indicate that Rebecca is more of an agent in the novel. Which name form is used, and why this particular name form is used, will lead too deep into the plot of the novel for this essay, and will be left for others to study. Instead, the following sections will investigate how the two women are presented in the text.

4.2 Keywords and Semantic Fields

The AntConc tool is used to automatically create a keyword list for the study. This keyword list contains all keywords in the *VF corpus*. A keyword is a word that appears relatively more frequently in a text than in a reference corpus (Mahlberg 2010: 20). In this case the *VictA Corpus* is the reference corpus used. The keywords can be said to represent the "aboutness" as they are specific to the topic(s) of the text (Fischer-Starcke 2009:496). Table 3 below shows a list of the top 20 keywords sorted by keyness factor. McIntyre (2011:167) defines keyness as

"the extent to which the frequency of lexical items in one corpus differs from their frequency in a larger reference corpus".

Rank	Frequency	Token
1	1045	crawley
2	628	osborne
3	638	amelia
4	540	rawdon
5	540	dobbin
6	511	rebecca
7	435	sedley
8	472	pitt
9	424	jos
10	421	becky
11	385	major
12	659	george
13	12963	and
14	252	briggs
15	200	steyne
16	194	emmy
17	168	bute
18	145	georgy
19	1052	miss
20	17507	the

Table 3. The initial keyword list of Vanity Fair. The tokens that are not names of characters in the novel are highlighted. Note that miss is ambiguous, as it can be both a verb and a form of address. (Running the concordance tool in case sensitive mode, shows that out of 1052 occurrences, only 11 refer to the verb form.)

Among the top 20 keywords listed in Table 3, no less than 16 are names of individuals in the text. In addition, two of the remaining four words are *major* and *[M]iss*, both of which can be used as a title or form of address. The last two words are the function words *and* and *the*. The high proportion of person names and titles at the top of the keyword list may be an indication that Thackeray's focus is on characters rather than e g locations or objects.

A new, modified keyword list is created by manually weeding out all proper names and function words from the initial keyword list. This is in part a subjective process which lends itself to the introduction of errors. Some of the terms in the initial list are

ambiguous, e g *vanity*, which is sometimes part of *Vanity Fair* – in itself ambiguous – and sometimes just a descriptive adjective. Since the majority of occurrences referred to *Vanity Fair*, the word was removed. Another ambiguous term is *Miss* (noun) vs *miss* (verb). Running the concordance tool in case sensitive mode showed that there were about 100 times as many nouns as verbs, and the word was kept in the keyword list. Table 4 below, shows a list of the top 50 keywords in the new list.

rank	token	rank	token	rank	token	rank	token
1	major	14	great	27	honour	40	military
2	miss	15	army	28	natured	41	marquis
3	captain	16	dear	29	woman	42	mustachios
4	old	17	young	30	drove	43	emperor
5	little	18	poor	31	spinster	44	dinner
6	colonel	19	gentleman	32	prince	45	british
7	regiment	20	city	33	boy	46	artless
8	ladies	21	honest	34	park	47	officer
9	mrs	22	famous	35	baronet	48	brother
10	carriage	23	lord	36	horses	49	women
11	french	24	queen	37	civilian	50	elephant
12	family	25	opera	38	wife		
13	friend	26	rectory	39	reverend		

Table 4. The top 50 keywords in the VF corpus after removal of proper names and function words.

As previous corpus stylistic studies have shown, analysing keywords often yields valuable information about a text (Fischer-Starcke 2010, Yazdanjoo et al 2016). One of the ways in which keywords may be ordered is by semantic field. The keywords from Table 4 above are grouped into eight semantic fields. The semantic fields are *Referring to men*, *Referring to women*, *Title or polite reference*, *Military vocabulary*, *Location*, *Descriptions (adjectives)*, *Familial or social relationship*, and *Material properties*. A keyword can belong to several semantic fields, e g *major* belongs to the fields *Referring to men*, *Title or polite reference*, and *Military vocabulary*. See Appendix A for a complete listing of the semantic fields and their constituents. Again, this is a subjective grouping, and both the

choice of fields, as well as the distribution of keywords could be discussed. Given the limited scope of this study, the impact of a dubious categorization will most likely be fairly inconsequential for the overall result, as it will not be based on these groupings per se. They will however play a role for the selection of keywords that will become the subjects of closer inspection in the following section. With a different definition of semantic fields or distribution of keywords, other keywords may have been chosen for closer study and the discussion in the next section of the essay would have dealt with another set of words. However, as long as the end result is valid, it may be of less importance in a smaller study such as this, which examples are used to highlight the characterisation of the protagonists in *Vanity Fair*.

4.2.1 Semantic Fields and the Female Protagonists

In order to test which of the semantic fields is most likely to have bearing on the two female protagonists, all of their name forms were run in the concordance tool, to see which of the keywords co-occurred with references to the two women. The contextual span was set from five words to the left of the name (L5) to five words to the right of the name (R5). The minimum collocating frequency was set to three. Table 5 below illustrates which keywords collocate with the names of the two women.

Collocates with Amelia's name forms only	Collocates with Rebecca's name forms only	Collocates with both women's name forms
boy	colonel	brother
dinner	family	captain
drove	honour	carriage
gentleman	horses	dear
regiment	lord	friend
wife	queen	great
women		ladies
		little
		major
		Miss
		Mrs
		old
		poor
		woman
		young

Table 5. Keywords in collocate positions to the female protagonists' name forms. The contextual span is L5 to R5. The minimum collocate frequency is 3.

It turns out that all of the keywords in the semantic field *Familial or social relationship* are present in Table 5; *Miss, Mrs, family, friend, boy, wife, brother*. This is the only semantic field for which all constituents are collocates with at least one of the name forms of the two women. In the novel, both protagonists become wives and mothers of a little boy. As the first obligation of the Victorian middle-class woman was to be a devoted wife and mother (Altick 1973:53), a closer look at the two keywords *wife* and *boy* may reveal something about how Thackeray chooses to present the two protagonists.

There are 288 instances of the keyword *boy* in the *VF Corpus.* Quite a few of these instances are part of the phrase *my boy*, which is and old-fashioned "friendly way of talking to a man" (*Cambridge Advanced Learner's Dictionary*, 2008). Sorting the concordance lines by left-hand collocations, yields eight instances containing the phrase *her boy*. These are likely to concern the sons of the two women, and so may be of interest to this study:

(1)	many by bursting into tears about	**her boy**	and exhibiting the most frantic grief when
(2)	she had given up everything for	**her boy**	; how she was careless of her parents in
(3)	She sees him, but he is not	**her boy**	any more. Why, he rides to see the
(4)	and was he not the father of	**her boy**	?) And as for the separation scene from the
(5)	her to change her mind respecting	**her boy**	. Her father had met with fresh misfortunes
(6)	very quick) when she should see	**her boy**	and how good and wise he had grown.
(7)	Here it was that she tended	**her boy**	and watched him through the many ills of
(8)	her parents, and given up	**her boy**	, when it seemed to her her duty to

The phrase *her boy* is a simple possessive phrase. One of the strengths of a concordance tool is its capability to find *all* occurrences of a word or a string of words in large text volumes and to show them in context (Mahlberg 2013: 8). This makes it possible to analyse a lengthy literary text at a level of detail which is nearly impossible to do with traditional close reading. Looking at the phrase *her boy* in context, shows that only lines (1) and (4) relate to Rebecca's boy, which in a way seems predictable to a reader of the novel, as Rebecca shows much less interest in her boy than Amelia does in her son. In addition, both of those instances are part of indirect recounts of Rebecca's (false) lamentations about the loss of her son and her husband ("the father of her boy") which further accentuates her emotional distance to her son. The other six lines refer to Amelia and her boy. In all of the instances, it is clear that Amelia's ties to her boy are strong and quite central to her character. In fact, looking at lines (2), (3), and (8) in context, there is a more or less openly stated view that Amelia's ties to her boy, and the way she sees him, are not well balanced, but that her life is too much focused on him. These ideas are not inferred by explicit linguistic patterns but are identified when the quantitative analysis is complemented by a qualitative analysis.

The second keyword to investigate further is *wife.* There are 288 instances of this word in *Vanity Fair*. Looking at the collocates immediately to the left (L1) of *wife* can show us which adjectives are used to qualify the noun in the novel: *admiring*, *astonished*, *beetle-browed*, *charming*, *dearest*, *deceased*, *faithful* (3 instances), *false* (2 instances), *first*, *good* (5 instances), *kind-hearted*, *late*, *little*, *old*, *poor*, *pretty*, *rebellious*, *Scotch*, *secluded*,

second, *tenth*, *thrifty*, *truest* and *young* (2 instances). A majority of these adjectives can be said to be neutral or positive, and so are well in line with the Victorian view of the woman as "The Angel in the House" (Altick 1973: 53). A closer look at the 37 instances of *wife* that have a qualifying adjective collocated in the L1-position reveals which adjectives are used to describe Rebecca's or Amelia's wifely qualities. The result of a review of the (adj)+*wife* phrases in context is presented in Table 6 :

	Adjectives qualifying *wife*
Referring to Amelia	charming, dearest, faithful, good, little (2), pretty, young
Referring to Rebecca	false (2), good, little (4), secluded, truest

Table 6. The adjectives used to qualify the keyword wife *when referring to one of the female protagonists.*

Table 6 indicates that Amelia is more often described as having the positive wifely qualities coveted by the Victorians than Rebecca is. Closer study of the concordance lines and the keywords in context reveals additional information about the attribution of the adjectives in Table 6 above. There are two instances when Rebecca is the one doing the attribution, which makes the reader see them in a different light:

(9) going with you to visit that foolish **little wife** of yours; as if I care a pin for either of
(10) man in the world. I was the **truest wife** that ever lived, though I married my husband

In line (9), Rebecca is talking condescendingly to Amelia's husband in a way which exposes him and his vanity to the reader, but at the same time belittles Amelia *and* makes the reader less sympathetic to Rebecca. In this instance, *little* cannot be said to have the positive connotation it generally has. In line (10), Rebecca is describing her relation to her late husband, in an attempt to snare another man. The reader knows that it is a lie, and so the utterance will infer something completely different about Rebecca's character than the lexical meaning of the phrase would imply.

The study of keywords is a way to look at the larger themes of a text, its "aboutness", as discussed earlier. It seems that approaching the text from this angle, could yield interesting information about the characters. In the case of Rebecca and Amelia, their roles as wives and mothers were highlighted. By quantitative measures, Amelia appeared to be the better wife and the more motherly of the two women. Adding a more qualitative analysis adjusted the picture somewhat, showing that beneath the surface, there were some negative attitudes towards Amelia hidden in the text. But this was also the case for Rebecca.

In the next section, the focus will be entirely on the *VF Corpus* and what can be gleaned from it, without the support of a reference corpus.

4.3 Collocates and Attributions

In order to get a more complete picture of how the two protagonists are characterized, it may be useful to investigate which adjectives in general are most frequently used for a direct description of the two women. Table 7 below lists the collocates immediately to the left (L1) of each character's different name forms which can be seen as descriptive of the two women. The most frequent collocating adjectives immediately to the left of one of the names, for both female protagonists, are first and foremost *little*, *dear*, and *poor*. These are adjectives that match the ideal Victorian woman; a fragile, dependent creature (Altick 1973, 53). It is worth noting that, despite being referred to by name less often than Rebecca in total in the novel, Amelia has a higher frequency of adjectives collocated with her name forms. In Table 7 words with a negative connotation are highlighted in red.

Rebecca		Amelia	
Collocate	**freq**	**Collocate**	**freq**
little	20	little	32
dear	18	dear/dearest	26
poor	7	poor	22
affectionate	2	pretty	2
accomplished	1	sweetest	2
admired	1	tender-hearted	2
amiable	1	amiable	1
beloved	1	astonished	1
darling	1	beloved	1
delighted	1	darling	1
enchanting	1	gentle	1
exclusive	1	imprudent	1
favourite	1	sentimental	1
gentle	1	wounded	1
good-humoured	1	young	1
grateful	1		
ingenious	1		
intoxicated	1		
odious	1		
scheming	1		
unfortunate	1		
unlucky	1		
unprotected	1		
virtuous	1		
Total	67		95
Normalized	0,55		0,98

Table 7. Adjectives collocated immediately to the left of one of the names used for the two women. Frequency of each collocate. The normalized number shows descriptive collocating adjective per 10 mentions of name.

It may be argued that *poor* has a negative connotation. Both girls, initially just Rebecca, but Amelia too after the demise of her father's business, are almost always poor in the sense of having very little money. In this novel however, the adjective *poor* immediately before one of the women's names is used in the sense that she is "deserving sympathy" (*Cambridge Advanced Learner's Dictionary*, 2008). A few sample concordance lines of *poor* can be used to illustrate this:

(11)	clear to every soul in the house, except	**poor**	Amelia, that Rebecca should take
(12)	Not that I dislike	**poor**	Amelia: who can dislike such a harmless
(13)	compare her with that	**poor**	Mrs. Osborne, who couldn't say boo to a goose.
(14)	Becky, of course, quite outshone	**poor**	Emmy, who remained very mute and timid

Judging by the frequency, it is Amelia who deserves most sympathy from the reader by far. She is attributed the word *poor* three times as often as Rebecca is in absolute numbers and almost six times as often in normalized numbers. Due to the repetition of these attributes readers may become more aware of them and use them for building their mental representation of the characters (Emmott & Alexander 2011:331).

When the numbers are normalized to show the frequency of collocated attributions in relation to the number of mentions of the protagonist by name, it is clear that Amelia is attributed a characteristic almost twice as often as Rebecca is, 0.98 vs 0.55 times per ten mentions by name (see Table 7). In addition, Thackeray uses a smaller set of adjectives to describe Amelia, and more consistently uses the top three (*little, dear/dearest, poor*) than he does for Rebecca. This harmonizes well with the reader's impression that Amelia's world is more confined and quiet than that of Rebecca and that Amelia herself is a less active character in the text; "as a 'domestic model' she becomes objectified in the suffocating silence of her childhood home" (Jadwin, 1992, 664). She is constant in her love for her no-good husband, even after his death, and the author is just as constant in the way he describes her. By attributing Rebecca with a larger variety of adjectives and with a wider spectrum of connotations, Thackeray is making her a rounder character than Amelia. The large variety of adjectives is also in line with Rebecca's behaviour, as she is constantly re-inventing herself to make the best of every new situation.

In the case of the adjective *poor* discussed previously, the normalized numbers show that the word is used not just three, but four times as often to describe Amelia as it is used to describe Rebecca. It could be interesting to also look at combinations of the three

most commonly used adjectives, i e when one is used to modify another. There are three combinations of these adjectives used in *Vanity Fair*; *dear little*, *poor little*, and *poor dear*.

In this case, the number of occurrences of the combined adjectives is relatively small, which makes it possible to also investigate all occurrences in context qualitatively, to identify who or what they are referring to, even when they are not collocated with one of the name forms. This can be illustrated by the following concordance lines, which are the first three to show up in AntConc for the phrase *dear little:*

(15)	our acquaintance, that she was a	**dear little**	creature; and a great mercy it is , both in life
(16)	very foolish vain fellow, and put my	**dear little**	girl into a very painful and awkward position
(17)	care for himself--not he; but his	**dear little**	girl should take the place in society to which,

In examples (15) – (17), the surrounding text is not enough to make it obvious who the *dear little* creature or girl is. However, the concordance tool makes it possible to jump straight from the concordance line to its location in the novel and there to read a longer passage of text to identify who *dear little* is referring to. The result of this exercise is described in Table 8 below.

	Total number of occurrences	Referring to **Amelia**	Referring to **Rebecca**	Referring to other subject
dear little	20	**11**	**2**	7
poor little	45	**25**	**10**	10
poor dear	20	**2**	**6**	12
total	85	**38**	**18**	

Table 8. Combinations of the three most common adjectives collocated immediately to the left of one of the names used for the two women. Note that the table shows <u>all</u> occurrences in the novel, not only those collocated with a name.

As Table 8 shows, in total 56 of the 85 adjective pairs in question which can be found in the novel refer to one of the two women (only 21 of these are collocated immediately to the left of a name.) Again, it is obvious that Amelia is the one who is most *dear*, *poor*, and *little*. She is attributed one of the adjective pairs more than twice as often as Rebecca is in absolute numbers (38 vs 18 instances). It is also interesting to note that no less than 5 out of

the 10 *poor little* attributed to Rebecca are done so by herself, typically when she is in a situation where she wants to invoke sympathy or downplay the threat she may be posing to her counterpart, which the following two sample lines illustrate:

(18) need not be jealous about me, my dear Miss Briggs. I am a **poor little** girl without any friends
(19) Think of that! **Poor little** me. I might have been Lady Crawley.

The statements in (18) and (19) would be examples of what Jedwin (1992) calls "female double-discourse", where "[a]mbitious women thus have little choice but to refashion "virtuous" discourse by reinflecting and exaggerating its rhetoric and concealing their unacceptable ideas beneath an acceptable surface" (Jedwin 1992, 664). It is also, in more linguistic terms, a way to infer character by way of presentation of speech which indirectly reveals the protagonist's motives and intentions. Example (18) can also be seen as a reply to the reaction of another character in the text. In contrast to these examples, Amelia does not describe herself in a similar manner, but all her attributions (of the investigated adjective pairs) are done either by other characters in the text or by the narrator.

The study of collocates is a text-internal approach, in the sense that the examples studied (*poor*, *little*, *dear*) were chosen not on the basis of a keyword list, but on what the primary corpus revealed when it was searched for collocates with the protagonists' names. Incidentally, the adjectives most used were also among the top keywords of the novel. By using these adjectives repeatedly to describe the two protagonists, and especially when describing Amelia, Thackeray manages to manifest these characteristics as part of the mental picture the reader is forming of her while reading the novel (Emmott & Alexander 2011:331).

5 Conclusions

The aim of this essay was to compare the way Thackeray chose to describe the two female protagonists in the novel *Vanity Fair* by using corpus linguistic and corpus stylistic tools and methods. With the help of the novel's keywords, eight semantic fields could be defined and

compared against the co-occurrence of keywords with the protagonists' names. This pointed to a semantic field describing *familial and social relationships* as having the strongest link with the protagonists. A closer study of two of the keywords in this field, *boy* and *wife*, indicated that Amelia's character is the one that is described more positively as a mother and wife, although she is not entirely faultless in these roles.

Looking at the different name forms used for the two women and their frequency in the text, showed that Rebecca is referred to by name more than 25% as often as Amelia is. This could be an indication of Rebecca being the more active character. On the other hand, Amelia's name is more often (almost twice as often) collocated with an adjective than Rebecca's is. However, Thackeray uses a much smaller variety of adjectives to describe Amelia than he does Rebecca. In terms of some of the theories of characterization (Stockwell & Mahlberg, Emmett & Alexander), this could be a way for the author to allow the reader to relatively quickly form a mental picture of Amelia, while allowing the picture of Rebecca to remain more fluid. The most frequently used adjectives for describing the two protagonists are *little*, *dear/dearest*, and *poor*, which is in line with the ideal picture of the Victorian woman.

Vanity Fair is a rich and complex novel, and this study has just begun to scratch the surface in terms of corpus stylistic studies. One interesting aspect of studying specific terms in context, was to find the semantic value of some instance of a word being offset by an apparent lie or misrepresentation in a speech act. With access to more advanced corpus tools that allow the tagging of speech, this could be an interesting area to study. But as mentioned previously, this is a rich text with a complex structure, so there are vast opportunities for further studies of the inner workings of the novel.

References

Altick, R.D. (1973). *Victorian people and ideas: a companion for the modern reader of Victorian literature*. New York: Norton.

Balossi, G. (2014). *A corpus linguistic approach to literary language and characterization: Virginia Woolf's The waves.* Amsterdam: John Benjamins Publishing Company, 2014.

Cambridge advanced learner's dictionary. (3. ed.) (2008). Cambridge: Cambridge University Press.

Cammack, Z. (2015). Amelia's Manual Manipulation of the Major in Thackeray's Vanity Fair, *The Explicator*, 73(2), 153-156

Carter, R. (2011). "Methodologies for Stylistic Analysis: Practices and Pedagogies". In McIntyre, D. & Busse, B. (red.) (2010). *Language and style: in honour of Mick Short*. Basingstoke: Palgrave Macmillan.

Cecil, D. (1934). *A Criticism of Life.* In Thackeray, W.M. (1994). *Vanity fair: an authoritative text, backgrounds and contents criticism*. New York: Norton.

Culpeper, J. (2000). A cognitive approach to characterization: Katherina in Shakespeare's *The Taming of the Shrew*. *Language and Literature,* Vol 9(4), 291-316.

Culpeper, J. (1996). Inferring character from texts: Attribution theory and foregrounding theory. *Poetics*, 23, 335-361.

Emmott, C. & Alexander, M. (2011). "Detective Fiction, Plot, Construction, and Reader Manipulation: Rhetorical Control and Cognitive Misdirection in Agatha Christie's *Sparkling Cyanide"* in McIntyre, D. & Busse, B. (red.) (2010). *Language and style: in honour of Mick Short*. Basingstoke: Palgrave Macmillan.

Fischer-Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's *Pride and Prejudice. International Journal of Corpus Linguistics*, 14, 492-523.

Fischer-Starcke, B. (2010). *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. London: Continuum.

Gilmour, R. (1982). *Thackeray: Vanity fair*. London: Edward Arnold.

Harden, E. F. (1967). "The Discipline and Significance of Form in *Vanity Fair*" in Thackeray, W.M. (1994). *Vanity fair: an authoritative text, backgrounds and contents criticism*. New York: Norton.

Heidbrink, H. (2011). "Fictional Characters in Literary and Media Studies" in Eder, J. Jannidis, Fotis, and Schneider, Ralf, eds. *Revisionen : Characters in Fictional Worlds : Understanding Imaginary Beings in Literature, Film, and Other Media*. Berlin/Boston, DE: De Gruyter, 2011.

Jadwin, L. (1992). "The Seductiveness of Female Duplicity in *Vanity Fair*". *Studies in English Literature, 1500-1900,* Vol 32, No. 4, Nineteenth Century (Autumn, 1992), pp. 663-687

Leech, G. (2011). "Analysing Literature through Language: Two Shakespearean Speeches" in McIntyre, D. & Busse, B. (red.) (2010). *Language and style: in honour of Mick Short*. Basingstoke: Palgrave Macmillan.

Litvak, J. (1996). Kiss Me, Stupid: Sophistication, Sexuality, and "Vanity Fair". *NOVEL: A Forum on Fiction*, Vol. 29 (2), 223-242

Louw, B. (1993). "Irony in the text or insincerity in the writer? – the diagnostic potential of semantic prosodies", in: Baker, M., G. Francis, E. Tognini-Bonelli (eds.) (1993), *Text and Technology*, Amsterdam: John Benjamins Publishing Company.

Mahlberg, M (2010). Corpus Linguistics and the Study of Nineteenth-Century Fiction. *Journal of Victorian Culture,*15, 292-298.

Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. New York: Routledge.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. London: Routledge.

McIntyre, D. (2011). “Dialogue and Characterization in Quentin Tarantino’s *Reservoir Dogs:* A Corpus Stylistic analysis” in McIntyre, D. & Busse, B. (red.) (2010). *Language and style: in honour of Mick Short*. Basingstoke: Palgrave Macmillan.

Pietka, R. (2010). Thackeray's Vanity Fair. *The Explicator*, 68:4, 239-241.

Stockwell, P. & Mahlberg, M. (2015). Mind-modelling with corpus stylistics in *David Copperfield*. *Language and Literature*, 24(2), 129-147.

Stubbs, M (2001). Text, Corpora, and Problems of Interpretation: A Response to Widdowson. *Applied Linguistics.*22/2, 149-172.

Thackeray, W.M. (1994). *Vanity fair: an authoritative text, backgrounds and contents criticism*. New York: Norton.

Van Peer, W. (ed.), 1989. The taming of the text: Explorations in language, literature and culture. London: Routledge.

Yazdanjoo M., Sabbagh, M., Shahriari H. (2016). Stylistic Features of Holden Caulfield’s Language in J. D. Salinger’s *The Catcher in the Rye*: A Corpus-Based Study. *English Studies*, 97(7), 763-778.

Corpus Material

Braddon, M.E. (1862). *Lady Audley’s Secret.* Retrieved from http://www.gutenberg.org/cache/epub/8954/pg8954.txt [October 2013]

Brontë, C. (1847/1897). *Jane Eyre.* London: Service & Paton. Retrieved from http://www.gutenberg.org/cache/epub/1260/pg1260.txt [October 2013]

Brontë, E. (1848/1910). *Wuthering Heights.* Retrieved from http://www.gutenberg.org/cache/epub/768/pg768.txt [October 2016]

Carroll, L. (1865). *Alice’s Adventures in Wonderland.* Retrived from http://www.gutenberg.org/cache/epub/11/pg11.txt [November 2013]

Collins, W. (1868). *The Moonstone.* Retrieved from http://www.gutenberg.org/cache/epub/155/pg155.txt [October 2013]

Dickens, C. (1860-61/1867). *Great Expectations.* Retrieved from http://www.gutenberg.org/cache/epub/1400/pg1400.txt [October 2013]

Disraeli, B. (1845) http://www.gutenberg.org/files/3760/3760-0.txt [October 2016]

Elliot, G. (1871-1872). *Middlemarch.* Retrieved from http://www.gutenberg.org/cache/epub/145/pg145.txt [October 2013]

Gaskell, E. (1854-1855). *North and South.* Retrieved from http://www.gutenberg.org/cache/epub/4276/pg4276.txt [October 2013]

Lawrence, G. A. (1857). *Guy Livingstone.* Retrieved from http://www.gutenberg.org/cache/epub/17084/pg17084.txt [November 2013]

Hughes, T. (1857). *Tom Brown's School Days.* Retrieved from http://www.gutenberg.org/files/32224/32224-0.txt [October 2016]

Thackeray, W.M. (1848). *Vanity Fair.* Retrieved from http://www.gutenberg.org/cache/epub/599/pg599.txt [October 2013]

Trollope, A. (1858). *Doctor Thorne.* Retrieved from http://www.gutenberg.org/cache/epub/3166/pg3166.txt [October 2013]

Wood, E. (1861). *East Lynne.* Retrieved from http://www.gutenberg.org/files/3322/3322-0.txt [November 2016]

Software

Anthony, L. (2014). AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Appendix A

Semantic field	Constituent keywords
Referring to men	Major, captain, colonel, gentleman, lord, prince, boy, baronet, reverend, marquis, emperor, officer, brother
Referring to women	Miss, ladies, Mrs, queen, woman, spinster, wife, women
Title or polite reference	Major, captain, colonel, gentleman, lord, prince, baronet, reverend, marquis, emperor, Miss, ladies, Mrs, queen
Military vocabulary	Major, captain, colonel, regiment, army, civilian, military, officer
Location	Regiment, city, opera, rectory, park
Descriptions (adjectives)	Old, little, French, great, dear, young, poor, honest, famous, (-natured), British, artless, honour
Familial or social relationship	Miss, Mrs, family, friend, boy, wife, brother
Material properties	Carriage, horses, mustachios

*The keyword list sorted into semantic fields. Some keywords show up in more than one semantic field, whereas three keywords (*honour, drove, elephant*) do not belong to any of the fields.*